Raport ştiinţific şi tehnic in extenso

Program: *Parteneriate în domenii prioritare*

Proiect

Titlul: *Sistem de Asistenta pentru Cladiri Inteligente, controlat prin Voce si Vorbire Naturala*

Acronim: *ANVSIB*

Data: *20.09.2015*

Etapa: *2/2015*

Activitatea / activităţile:

- *Activitatea 2.1 Proiectarea corpusului de vorbire (partea a II-a)*
- *Activitatea 2.2 Inregistrarea corpusului de vorbire*
- *Activitatea 2.3 Segmentarea si adnotarea corpusului de vorbire (partea I)*
- *Activitatea 2.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii (partea a II-a)*
- *Activitatea 2.5. Implementarea unui controller de voce (partea I)*
- *Activitatea 2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de camera) (partea a II-a)*
- *Activitatea 2.7. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)*

Număr contract: *32 / 2014*

Acord de colaborare: *10677/24.06.2014 UPB, 1744/03.06.2014 IWAVE, 164/19.06.2014 RACAI*

Autoritatea contractantă: *Unitatea Executivă pentru Finanţarea Învăţământului Superior, a Cercetării, Dezvoltării şi Inovării*

Conducător de proiect: *UPB*

Director de Proiect: Horia Cucu

Cuprins

1 Obiectivele activităților desfăşurate

În a doua faza a proiectului ANVSIB, obiectivele RACAI au fost finalizarea proiectării corpusului de vorbire (1), înregistrarea lui (2) cat si segmentarea si adnotarea (3) pentru a produce un model acustic de calitate.

Obiectivul activitatii 2.4 a fost crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii..

Obiectivul acivitatii 2.5 a fost implementarea unui controller de voce.

Obiectivul activității 2.6 a fost achiziționarea unei baze de date de vorbire in condiții reale.

Obiectivul activității 2.7 a fost implementarea parametrilor PNCC (Power Normalized Cepstral Coefficients) în sistemul de recunoaştere a vorbirii al UPB şi re-evaluarea performanțelor sistemului după această modificare.

2 Perioada de desfăşurare şi personalul implicat

Nr.	Denumire activitate	Nume si prenume	Perioada
1.	*Activitatea 2.1 Proiectarea corpusului de vorbire (partea a II-a)*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.01.2015 – 31.03.2015*
2.	*Activitatea 2.2 Inregistrarea corpusului de vorbire*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.01.2015 – 31.05.2015*
3.	*Activitatea 2.3 Segmentarea si adnotarea corpusului de vorbire (partea I)*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.01.2015 – 31.08.2015*
4.	*Activitatea 2.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii (partea a II-a)*	*Badulescu Marian*	*februarie 2015*
5.	*Activitatea 2.5. Implementarea unui controller de voce (partea I)*	*Badulescu Marian*	*februarie 2015*
6.	*Activitatea 2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de camera) (partea a II-a)*	*Cucu Horia; Buzo Andi; Marinescu Radu; Dogariu Mihai; Minca Florentina*	*01.01.2015 – 30.04.2015*
7.	*Activitatea 2.7. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)*	*Buzo Andi; Petrică Lucian; Banică Alina*	*01.05.2015 – 30.09.2015*

3 Rezumat activități

3.1 A2.1 Proiectarea corpusului de vorbire (partea a II-a)

Legat de proiectarea corpusului am obtinut textele initiale (extrase din Wikipedia). A urmat faza de dezvoltare a uneltei de curatare corpus prin care am eliminat toate propozitiile care nu corespundeau cu standardul nostru. Mai departe, am dezvoltat o unealta care primeste ca intrare o serie de propozitii si ofera ca iesire un subset sortat de propozitii care sunt balansate fonetic. Astfel, am obtinut cateva mii de propozitii curate si balansate fonetic (la nivel de trifonemi). Asadar, am finalizat proiectarea corpusului de vorbire.

3.2 A2.2 Inregistrarea corpusului de vorbire

In continuare, activitatea de inregistrare s-a efectuat cu ajutorul a doi actori subcontractati (voce masculina si feminina). Inregistrarile s-au efectuat sub supravegherea noastra pentru a asigura calitatea corpusului. Astfel, s-au facut re-inregistrari unde a fost nevoie pentru a corecta linia prozodica a actorului. De asemenea, eventualele greseli de gramatica, inconsistente si alte erori din text au fost corectate fie direct la inregistrare fie ulterior inregistrarilor. De exemplu, cuvintele dintr-o limba straina care au patruns totusi in propozitii si care au fost inregistrate, au fost introduse manual in dictionarul de transcriere fonetica generat de noi. In total s-au inregistrat aproximativ 20 de ore de voce de inalta calitate.

3.3 A2.3 Segmentarea si adnotarea corpusului de vorbire (partea I)

Ultima activitate desfasurata de RACAI a constat in adnotarea corpusului inregistrat. Astfel, s-a facut segmentarea propozitiilor folosind o unealta dezvoltata intern, rezultand o serie de propozitii corelate cu textul procesat si neprocesat al fiecareia. Pentru a adnota aceste propozitii a fost necesara adaptarea (si uneori dezvoltarea) uneltelor RACAI de procesare de limbaj natural. Am folosit unelte de segmentare in propozitii si in cuvinte, de adnotare a partilor de vorbire, de silabificare, predictie de accent lexical, transcriere fonetica, etc. Odata adnotate cu toate aceste utilitare, propozitiile au fost introduse intr-o coada de procesare pentru a obtine modelele acustice necesare modulului de sinteza de voce. Descrierea aplicatiei de segmentare cat si procesul de adnotare si creare a modelului acustic sunt prezentate in capitolele urmatoare.

3.4 A2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de camera) (partea a II-a)

În cadrul activității 2.6 (Achiziționarea unei baze de date de vorbire in condiții reale (partea a II-a)) a fost creată o bază de date de înregistrări audio de propoziții în condiții reale, într-o casă inteligentă. Frazele înregistrate reprezintă:

- seturi de comenzi pe care utilizatorul le poate da casei inteligente
- monologuri ale unui vorbitor pe un subiect prestabilit
- dialoguri între mai mulți vorbitori
- episoade de tuse voluntară

În prima parte a acestei activități (Activitatea 1.5 din anul 2014) au fost realizate următoarele

- întocmirea unei liste cu toate echipamentele din casă ce pot fi automatizate
- stabilirea acțiunilor ce pot fi desfășurate cu aceste echipamente
- compunerea unei liste de comenzi cât mai cuprinzătoare
- stabilirea liste finale de comenzi luând în calcul o varietate de moduri de exprimare posibile

Continuând în etapa curentă, în cadrul activității 2.6 din anul curent au fost realizate următoarele:

- întocmirea unei liste cu subiecte pentru monologuri şi dialoguri
- întocmirea mai multor liste cu comenzi frecvente pentru a fi înregistrate de diverşi vorbitori
- redactarea unui document de consimțământ pentru utilizarea înregistrărilor
- contactarea şi instruirea persoanelor în vederea realizării înregistrărilor
- supravegherea şi îndrumarea persoanelor în vederea realizării înregistrărilor
- înregistrarea propriu zisă a clipurilor audio într-o casă inteligentă
- post-procesarea bazei de date de înregistrări audio

Mulțumim pe această cale Laboratorului de Informatică din Grenoble care ne-a pus la dispoziție casa inteligentă în care s-au realizat înregistrările audio şi ne-au sprijinit pe tot parcursul activității cu materiale, sfaturi, sugestii, acces la persoane vorbitoare de limba română din Grenoble şi altele.

Activitatea a fost realizată în proporție de 100%.

3.5 A2.7. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)

În cadrul activității 2.7 (Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)) a fost realizat un modul de extragere a parametrilor vocali PNCC (Power Normalized Cepstral Coefficients) care folosesc a) analiza de putere pe termen mediu, în care parametrii de mediu sunt estimați pe o durată de timp mai mare decât cea utilizată în mod frecvent pentru procesarea vorbirii, şi b) frecvența de uniformizare.

Am contribuit la implemnetarea acestor parametrii robuşti la zgomot în CMU Sphinx, un toolkit open-source de recunoaştere a vorbirii, care este nucleul sistemului de recunoaştere a vorbirii al laboratorului de cercetare SpeeD din cadrul UPB. Apoi, am experimentat cu aceşti noi parametri ca înlocuitori pentru parametrii MFCC (Mel Frequency Cepstral Coefficients) tradiționali şi am obținut rezultate mai bune pentru toate tipurile de vorbire zgomotoasă pe care am încercat să o transcriem.

Activitatea a fost realizată în proporție de 100%.

4 Descrierea ştiinţifica şi tehnică, cu punerea în evidenţă a rezultatelor activităţilor si gradul de realizare a obiectivelor

4.1 Introducere

Tehnologii din cadrul procesarii limbajului, cum ar fi recunoaşterea automată a vorbirii (ASR) şi sinteza textului (TTS) sunt pietrele de temelie ale oricarei interfeţe în limbaj natural. În conformitate cu seria MetaNet White Papers (Rehm si Uszkoreit, 2012), prin analiza stadiului actual al resurselor si instrumentelor disponibile pentru prelucrarea vorbirii, limba română a fost clasificat în clasa de suport fragmentar, împreună cu alte 14 limbi europene (stadiul `fragmentar` fiind al doilea cel mai mic grad din cinci). Ca atare, importanţa unor resurse semnificative TTS este intrinsecă obiectivului de a evita disparitia digitala a limbii române.

Sinteza textului (TTS) necesită un corpus proiectat şi înregistrat cu mare atentie pentru a produce rezultate de înaltă calitate. La proiectarea unui astfel de corpus trebuie să se ia în considerare un număr mare de variabile, cum ar fi: (a) faptul că datele înregistrate trebuie să ofere o acoperire bună in limba tinta si pe domeniul dorit, (b) domeniul din care propoziţiile provin va avea un impact mare asupra prozodiei care va fi "învăţata" de către algoritmii de învăţare automata care stau la baza sistemului TTS, (c) naturaletea vocii vorbitorului va fi un factor decisiv in a fi acceptata vocea de cătreascultători sau nu, si (d) vorbitorul (cel care inregistreaza corpusul) trebuie să înţeleagă limitele care guvernează capacitatea de modelare metrica a sistemului TTS în sine şi trebuie să isi limiteze (compromis) o parte din expresivitatea vocii sale, în scopul de a reduce numărul de artefacte prozodice rare, care ar putea fi introduse în caz contrar în corpus.

În plus, trebuie să se constate că, indiferent de sistemul de sinteză de vorbire folosit, fie bazat pe metoda concatenativă, fie sinteză parametrica, ambele metode produc rezultate mai bune atunci când sunt furnizate cu corpus de înaltă calitate, care necesita utilizarea dispozitivelor de înregistrare profesionala şi de instrumente, de preferinţă în condiţii de studio. De asemenea, sistemele de sinteză a vorbirii bazate pe metoda concatenativă sau pe sinteză parametrică statistică cunosc o îmbunătăţire a rezultatelor care depind foarte mult de volumul datelor de intrare. Elementele prozodice învăţate automat din corpus influenţează modul de folosire a sistemului de sinteză a vorbirii. Astfel, un sistem de sinteză ce foloseşte un corpus din domeniul ştiri, nu este uşor scalabil pentru a fi utilizat în citirea de nuvele sau basme. La fel de importantă este distribuţia datelor de intrare: unităţile acustice de bază folosite în procesul de antrenare şi sinteză trebuie să apară cel puţin o dată, iar observarea lor în mai multe contexte ajută la generarea unui model statistic şi acustic mai bine adaptat din punct de vedere prozodic. Asadar, în procesul de proiectare a unui corpus de voce destinat a fi folosit ca date de antrenare pentru un sistem de sinteză a vorbirii, trebuie să ţină cont de următoarele cerinţe: (a) corpusul trebuie să acopere mai multe domenii pentru a putea genera şi utiliza, în funcţie de caz, un model corespunzător cu domeniul textului ce urmează a fi sintetizat; (b) unităţile acustice trebuie să ofere o acoperire completă a inventarului fonetic al unei limbi; (c) este preferabil ca fiecare unitate să apară de mai multe ori în contexte diferite.

În următoarele secţiuni vom descrie procesul de selectare a propoziţiilor care vor face parte din corpus, procesul de înregistrare si in final de adnotare a lui.

4.2 A2.1 Proiectarea corpusului de vorbire (partea a II-a)

Dupa cum a fost mentionat anterior, una din conditiile necesare pentru a dezvolta un sistem TTS este ca un corpus trebuie sa ofere o buna acoperire pe limba dorita si pe domeniul specific dorit. Cu alte cuvinte, (a) corpusul trebuie sa fie balansat fonetic din punct de vedere al unitatilor de vorbire (fonemi, difonemi, etc.).

Textul sursa pentru corpus este extras din Wikipedia si contine un numar de propozitii care au fost selectate prin folosirea algoritmului <greedy> (vom reveni asupra acestuia mai tarziu) pentru a asigura completitudinea domeniului fonetic al limbii romane. Propozitiile sunt tratate ca unitati de sine statatoare (contextul nu este necesar), astfel incat vorbitorul trebuie sa inregistreze propozitiile avand o linie prozodica moderata si este obligat sa-si limiteze interprearea narativa la propozitia in sine.Deoarece textul extras din Wikipedia contine numeroase erori si este departe de a fi un text clar si usor de citit, trebuie sa folosim un numar de reguli euristice pentru a elimina sau/ori pentru a corecta propozitiile. Mai jos se regasesc enumerati pasi procesati si aplicati:

a. Corpusul trebuie impartit in propozitii si pe urma in cuvinte, pastrand doar liniile care nu au mai mult de 20 de cuvinte. Folosind unealta de despartire in propozitii dezvoltata in cadrul Racai (bazat pe un algoritm de tip Maximum Entropy), am obtinut peste 5 milioane de astfel de propozitii.
b. Vor fi eliminate toate spatiile dinainte si dupa propozitie sau caracterele ce nu pot fi printate.
c. Vor fi eliminate toate propozitiile care contin urmatoarele caractere: ./ ● ¾ ’ ◦,paranteze, bare oblice, citate, etc (mai multe caractere vor fi adaugate manual), dar si propozitiile care contin abrevieri sau cuvinte precum: Sos., Cal., .ro., uk., www., etc. Toate aceste reguli au fost adaugate manual deoarece exista un numar mare de propozitii care contin aceste cuvinte si care nu sunt potrivite pentru inregistrari.

   Unele reguli sunt bazate pe expresii regulate, ca de exemplu un cuvant care are caractere latine de la A la Z, pe cand altele sunt niste simple conditii , ca de exemplu o propozitie sa nu aiba un anume subsir.
d. Vor fi eliminate toate propozitiile care contin numere.
e. Vor fi eliminate toate propozitiile care sunt scrise cu litere mari (in general titlurile)
f. Vor fi eliminate toate propozitiile cu mai putin de trei cuvinte cu urmatoarele exceptii: daca propozitia contine un singur cuvant, atunci acesta ar trebui sa se regaseasca in lexiconul limbii romane. Daca propozitia contine doua cuvinte, atunci ambele ar trebui sa se regaseasca in lexicon, pentru ca propozitia sa nu fie stearsa. Daca propozitia contine trei cuvinte, atunci cel putin doua din cele trei cuvinte trebuie sa se regaseasca in lexicon (cuvinte romanesti, nu nume proprii precum numele persoanelor, locurilor, entitati denumite, etc.).
g. Vor fi eliminate toate propozitiile in care nu se regasesc cel putin 90% din cuvinte in lexicon (exceptia o vor face numele proprii). Aceasta regula a permis stergerea a numeroase propozitii care contineau cuvinte in limbi straine (chiar daca s-a folosit corpusul extras din Wikipedia romaneasca, tot am gasit un numar mare de propozitii care contin cuvinte in alte limbi).
h. Vor fi eliminate toate propozitiile in care nu se regasesc cel putin 90% din cuvinte scrise cu diacritice. Exista foarte multe propozitii in articolele din Wikipedia care nu au diacritice sau in care doar cateva cuvinte contin diacritice. Luam in considerare doar cuvintele pe care le gasim in lexicon si care ar trebui sa contina diacritice.
i. Se vor corecta cuvintele romanesti care obisnuiau sa se scrie diferit: i-(î) in forma corecta actuala a-(â). De exemplu, forma veche a cuvantului “cîte” a fost corectata conform noii scrieri la “câte”. Acest proces se realizeaza conform urmatorilor pasi:
   a. Prin folosirea lexiconului pentru a afla direct forma corecta a cuvantului: se creeaza o noua mapare intre lexicon (fiecare cuvant in parte) si forma ocurenta scrisa cu î din i in loc de â din a. Se va cauta in maparea predefinita a lexiconului cuvantul scris cu î din i si se va returna forma cuvantului scrisa cu â din a. Nu exista nicio coliziune intre cuvinte.
   b. Daca o astfel de forma ocurenta nu este gasita, atunci ne vom intoarce la regula algoritmului de baza: de fiecare data cand un cuvant cu un î din i este gasit, se scoate prefixul, se inlocuiesc toate aparitiile lui î din i cu â din a,se recompune cuvantul. Acest proces se supune unui set de reguli si exceptii din limba romana. Ne asteptam la o rata de corectie de aproape 100% a algoritmului.

Pas cu pas, numarul propozitiilor a scazut la aproximativ 252000 (doar aproximativ 19% dintre propozitii au trecut de faza de curatare si corectie). In mod interesant, majoritatea liniilor care au fost eliminate au continut numere (regula d) sau nu au continut un procent minim de cuvinte din lexicon (regula g). Pe setul astfel rezultat de propozitii, am aplicat un algoritm de balansare fonetica.

Pentru a mentine numarul de trifonemi individuali cat mai balansat (o balansare perfecta nu este posibila deoarece exista o serie de trifonemi foarte rari), am aplicat urmatorul algoritm:

1. Am generat un table de frecventa al trifonemilor pe intreg corpusul;
2. Daca o propozitie contine un trifonem rar (cu frecventa sub 100), se pastreaza;
3. Daca o propozitie contine doar trifonemi frecventi (cu frecventa peste media distributiti initiale), se elimina;
4. Actiunea predefinita: pastreaza propozitia;
5. In final, sorteaza propozitiile in ordinea ascendenta a frecventei trifonemilor: astfel se asigura un corpus balansat de la inceput, indifferent cate propozitii sunt inregistrate din corpus.

4.3 A2.2 Inregistrarea corpusului de vorbire

Pentru înregistrarea corpusului de voce am subcontractat serviciile a doi actori profesionisti obtinand o voce masculina si una feminina. Actorii au primit un instructaj despre scopul proiectului si caracteristicile necesare pentru a face astfel de inregistrari (stil, intonatie, linie prozodica, etc). De asemenea, ei au fost supravegheati partial in timpul inregistrarilor pentru a obtine calitatea de corpus de vorbire dorita.

Corpusul contine propozitii din Wikipedia. Acest text are ca durata de voce aproximativ 10 ore. In total, pentru cei doi actori, textul extras din Wikipedia insumeaza aproximativ 20 ore de înregistrări. Textele din Wikipedia contin erori, asadar, in timpul inregistrarilor, actorii au notat inconsistente in propoziții (erori gramaticale, de sens, cuvinte scrise gresit, fara diacritice, etc), ei citind corect textul, rămânând ca propozițiile sa fie corectata ulterior (ca text scris). De asemenea, au fost marcate cuvintele care nu apartin limbii romane (exista dese referiri la nume/locatii in engleza, titluri de piese muzicale, de organizatii, etc. pentru care un proces automat de transcriere fonetica ar genera erori). Aceste cuvinte au fost luate manual si introduse in lexiconul necesar procesului de transcriere fonetica, ținând cont inclusiv de modul in care a fost pronunțat fiecare cuvânt de fiecare actor.

In total, pentru fiecare ora de voce, actorii au petrecut aproximativ 3 ore pentru a inregistra si reinregistra textele, necesitand astfel un efort aproximativ de 60 de ore de inregistrare. Alaturi de efortul actorilor, pentru prima parte de inregistrari, unul din angajatii RACAI a participat alaturi de actori pentru a-i ghida, efort insumat la peste 18 ore.

Rezultatul acestei activități este reprezentat de pachete de fisiere tip .wav (necomprimate), la 44KHz, 2 canale (stereo). Pachetele contin mai multe propozitii inregistrare unele dupa altele, cu o pauza de cateva secunde intre ele. Deşi aceasta metoda necesita un pas aditional de procesare (fiind nevoie de o spargere ulterioara in propozitii individuale), a fost totusi necesara deoarece altfel efortul de inregistrare ar fi fost mult crescut.

4.4 A2.3 Segmentarea si adnotarea corpusului de vorbire (partea I)

Segmentarea corpusului de vorbire este o activitate necesara inainte de adnotarea efectiva. Segmentarea presupune prelucrarea fisierelor audio care contin mai multe propozitii inregistrate una după alta in mai multe fisiere audio care contin propozitii individuale. Simiar, textul este stocat in calupuri mari de propozitii si trebuie impartit a.i. sa corespunda cu fisierele audio in mod 1-la-1.

Aplicatia de segmentare este folosita pentru a usura adnotarea si segmentarea fisierelor audio, in acelasi timp corectand fisierele text sursa. Astfel ca intrare, aceasta aplicatie foloseste un fisier audio integral (nu este nevoie de fisiere mai mici) si un fisier text care reprezinta transcrierea audio. De preferat este ca fisierul text sa fie impartit in propozitii inainte de utilizare. Mai specific, este dorita adnotarea fiecarei propozitii cu timp start si timp stop pentru a putea apoi taia automat din fisierul audio care contine tot semnalul audio exact propozitia care se doreste.

Aplicatia are forma unei platfome web, fiind disponibila online. Este scrisa in PHP + Javascript. Baza de date folosita este de tip SQLite. Utilizatorii platformei se autentifica (nume de utilizator/parola), apoi apasa pe „Document nou”. Daca au deja un document in lucru, acest document va fi automat afisat pe ecran. Documentul va fi afisat atat sub forma audio (waveform) in partea de sus a ecranului, cat si sub forma text in partea de jos. Partea de afisare audio ca waveform ajuta utilizatorul sa intuiasca inceputul si finalul fiecarei propozitii (acolo unde este liniste in audio banda rosie are amplitudine mica – este ingusta). In timpul ascultarii fisierului audio exista o linie distincta care se deplaseaza deasupra semnalului indicand pozitia curenta in audio. In partea de jos a ecranului se afla propozitiile citite din baza de date. Procedeul de lucru este urmatorul: Utilizatorul apasa spatiu cand considera ca fiecare propozitie se termina, si spatiu pentru a incepe redarea propozitiei urmatoare. In lista de propozitii timpul start/stop se completeaza automat. Acolo unde este nevoie, utilizatorul editeaza, adauga sau sterge propozitii, astfel incat sa fie o corespondenta 1-la-1 intre ceea ce se aude cu propozitiile transcrise.

Odata cu terminarea propozitiilor din cadrul unui document se apasa pe „finalizare editare si deschidere document nou”. Acest buton confirma salvarea timpilor si modificarilor propozitiilor in baza de date si va trece la urmatorul fisier audio si document (text) neadnotat.

Deasemena, la apasarea butonului „Salveaza” toate modificarile curente vor fi incarcate in baza de date corespunzatoare documentului curent. Astfel, in orice moment un utilizator poate salva datele, inchide si relua lucrul la un timp ulterior fara a pierde nimic.

1. Odata incheiat procedeul de adnotare al fiecare propozitii cu timpul start/stop pentru toate documentele pentru care se doreste acest lucru, administratorul de sistem va rula un script numit „dump_corpus.php”. Acest script efectueaza urmatorii pasi:
    a. Creaza o lista cu toate documentele care au fost finalizate.
    b. Pentru fiecare document in parte, citeste din baza de date „sentences.db” timpii de start si stop ai fiecarei propozitii.
    c. Pentru fiecare propozitie, scriptul va crea 3 fisiere:
        i. Fisierul wav: acest fisier este decupat automat din fisierul principal, incepand de la timpul de start pana la timpul de final al propozitiei curente. Fisierul va fi denumit automat cu un indicator de nume unic si un numar de propozitie.
        ii. Fisierul txt: fisierul va avea acelasi nume ca fisierul wav, insa cu terminatia .txt. Acest fisier contine pe o singura linie propozitia nemodificata. Singura corectura care i se poate aduce este modificarea diacriticelor ş si ţ din varianta veche cu sedila in varianta noua cu virgula.
        iii. Fisierul lab: similar cu fisierul text, insa cu extensia „.lab” contine propozitia pe un singur rand, insa modificata. Toate literele sunt trecute in litere mici; Toate semnele de punctuatie sunt eliminate; Toate liniutele sunt eliminate si cuvintele care erau unite prin aceste linuite sunt unite la propriu (ex: „Ducandu-i” devine „ducandui”, sau „şi-o” devine „şio”).

Aceste fisiere sunt necesare mai departe pentru antrenarea modelelor fonetice. Aplicatia de segmentare este utila pentru a reduce volumul de munca necesar pentru a efectua operatiuni de taiere audio si corectare text. Utilizand aceasta platforma nu mai este necesara folosirea unei intregi suite de aplicatii neconectate intre ele (ex: editare audio cu Audacity, apoi editare text in notepad, apoi taiere audio manuala cu Audacity sau sox, etc).

Adnotarile cuvintelor inregistrate sunt automat generate de uneltele de procesare de limbaj natural ale RACAI. Astfel, avem unelte pentru: i) impartirea in propozitii si in cuvinte, ii) Chunking, iii) Adnotarea partilor de vorbire, iv) Silabificare, v) Prezicerea accentului lexical, vi) Transcrierea fonetica.

Am investit un efort semnificativ pentru a dezvolta aceste unelte in conditiile in care limba romana este o limba relativ dificila (contine multe forme flexionare). Am incercat sa punem la punct o platforma comuna (un algoritm comun) pentru uneltele de mai sus. Totusi, folosind algoritmi de tip data-driven (foarte utili in cea mai mare parte a cazurilor) nu am reusit sa dezvoltam o metoda mai buna pentru predictia prozodiei. Astfel sistemul nostru se bazeaza, ca aproape toate celelalte sisteme de predictie prozodica, pe informatii extrase doar din contextul local (cuvintele in sine). Totusi, vom continua lucrul in aceasta directie deoarece un sistem de predictie prozodica este esential pentru a avea un sistem de sinteza de voce performant / care sa sune natural.

Deoarece marea parte a uneltelor au fost prezentate deja in articole si publicatii anterioare, ne vom concentra doar pe o descriere sumara a lor.

Adnotarea partilor de vorbire este o tehnica utilizata in multe aplicatii NLP precum Regasirea de Informatii, Extragerea de Informatii, Traducerea Automata, Dezambiguizarea de Sens, Generarea Arborilor Sintactici, etc. Adontarea partilor de vorbire este necesara inclusiv in sinteza de voce deoarece de ea depinde dezambiguizarea omografelor si predictia prozodiei. Marea parte a utilitarelor de adnotare a partilor de vorbire folosesc algoritmi de tipul Hidden Markov Models (HMMs). Desemenea, sunt folositi algoritmi de tip Maximum Entropy (Berger et al., 1996; Ratnaparkhi, 1996), Retele Bayesiene (Samuelsson, 1993), Retele neuronale (Marques and Lopes, 1996), Conditional Random Fields (CRF) (Lafferty et al., 2001), etc. Desi aceste metode sunt bine vazute si des utilizate pe limbi de tipul Englezei, pe limbi flexionate (ca limba Romana) utilizarea lor aduce o serie de probleme din cauza lipsei datelor (asa numita „data-sparseness”). De exemplu, limba romana foloseste in jur de 1200 de adnotari de parte de vorbire, care pot fi reduse la 614 printr-un proces de compresie fara pierdere de date. Limba Ceha foloseste peste 2000 de astfel de adnotari, pe cand in limba Engleza nu sunt necesare mai mult de aproximativ 100. Au fost propuse diverse metodologii pentru a rezolva aceasta problema (e.g. Tiered Tagging (Tufiş, 1999)), in principal reducand lipsa de date statistice prin efectuarea mai multor cicluri succesive de adnotare, cu nivele de informatie progresiva. Noi folosim o unealta de andotare de parti de vorbire bazata pe retele neuronale, descrisa in (Boroş et al., 2013a). Ideea principala este de a genera o codare unica pentru fiecare parte de vorbire (tag)

astfel incat fiecare atribut morfologic sa primeasca un identificator unic. Fiecare tag este convertit intr-un vector de valori reale bazate pe identificatorul atributelor morfologice. O retea neuronala este antrenata sa gaseasca corelatii intre valori (ex: gen, caz, etc.) bazat pe un context de 5 cuvinte/taguri (2 taguri anterioare precum si urmatoarele 3 taguri probabile, prezise de un algoritm de tip Maximum Likelihood Estimation)

Aceasta metoda de adnotare are rezultate bune pe limba romana (98% acuratete). Principalele avantaje folosind aceasta tehnica sunt ca (1) prin controlul topologiei retelei, putem ajusta rapid balansul intre viteza de predictie si acuratete, si (2) permite mascarea atributelor fara a fi nevoie de redescrierea intregului set de tag-uri (de parti de vorbire). Astfel, sistemul poate fi reantrenat pentru diverse taskuri de procesare de limbaj natural unde nu este nevoie de predictia intregului set de taguri.

Restul de pasi de procesare in cadrul sintezei de voce sunt bazati pe algoritmul MIRA : Margin Infused Relaxed Algorithm (MIRA)(Cramer and Singer, 2003). Folosind acest algoritm pentru toate uneltele de procesare aplicate ulterior, obtinem un sistem unitar, usor de modificat. Algoritmul este descris in (Boroş et al., 2013b). Metoda folosita este cea onset-nucleus-coda (ONC), propusa in (Bartlett et al., 2008).

Datele astfel obtinute vor fi folosite pentru a produce un context prozodic, in care fonemii individuali vor avea urmatoarea structura: 1. Contextul fonetic derivat dintr-o fereastra de 5 fonemi (centrata pe fonemul curent), 2. Silaba curenta, 3. Partea de vorbire a cuvantului curent, 4. Pozitia fonemului in cadrul cuvantului, 5. Pozitia fonemului in cadrul silabei, 6. Pozitia silabei in cadrul cuvantului, 7.Pozitia relativa a silabei intre ultima silaba accentuata si urmatoarea, 8. Existenta accentului lexical pe silaba curenta, 9. Chunk-ul curent.

```
Acesta    DMSR      a-'ces-ta a ch e s t a
este      V3        'es-te    j e s t e
un        TSR       un        u n               Np1
simplu    ASN       'sim-plu  s i m p l         Np1
test      NSN       test      t e s t           Np1
.         PERIOD    .         pau
```

Figura 1 Adnotarea NLP a propoziţiei"Acesta este un simplu test"

```
   PPP:_/PP:_/CP:a/NP:ch/NNP:/e/SYL:a/NSYL:3/NPHON:6/SIW:start/SYLI:0/PI:0/STRESS:0/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/
WIC:/SENTTYPE:PERIOD
   PPP:_/PP:a/CP:ch/NP:e/NNP:/s/SYL:ces/NSYL:3/NPHON:6/SIW:mid/SYLI:1/PI:1/STRESS:0/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/
WIC:/SENTTYPE:PERIOD
   PPP:a/PP:ch/CP:e/NP:s/NNP:/t/SYL:ces/NSYL:3/NPHON:6/SIW:mid/SYLI:1/PI:1/STRESS:0/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/
WIC:/SENTTYPE:PERIOD
   PPP:ch/PP:e/CP:s/NP:t/NNP:/a/SYL:ces/NSYL:3/NPHON:6/SIW:mid/SYLI:1/PI:1/STRESS:0/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/
WIC:/SENTTYPE:PERIOD
   PPP:e/PP:s/CP:t/NP:a/NNP:/j/SYL:ta/NSYL:3/NPHON:6/SIW:end/SYLI:2/PI:2/STRESS:1/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/WI
C:/SENTTYPE:PERIOD
   PPP:s/PP:t/CP:a/NP:j/NNP:/e/SYL:ta/NSYL:3/NPHON:6/SIW:end/SYLI:2/PI:2/STRESS:1/POS:DMSR/WISENT:start/WISENTS:0/WISENTE:5/CHUNK:/WI
C:/SENTTYPE:PERIOD
```

Figura 2 Contextul prozodic al cuvantului "Acesta" din exemplul anterior

- Pozitia cuvantului in chunk-ul curent
- Tipul propozitiei
- Lungimea propozitiei
- Lungimea cuvantului

Alinierea la nivel de fonem este ultima parte necesara procesului de andotare. Acest lucru l-am facut folosind utilitarul numit HMM Toolkit (HTK) (Young et al., 1993), urmarind urmatoarea procedura:

1. Generarea transcrierii fonetice (folosind HDMan) pentru toate cuvintele din corpus folosind o varianta imbunatatita de dictionar RSS (Stan et al., 2011). Cuvintele negasite in dictionar care sunt gasite in corpus au fost automat transcrise folosind trei agloritmi de transcriere fonetica: algoritmul anterior descris bazat pe MIRA, un algoritm bazat pe Maximum Entropy si un al treilea algoritm denumit DLOPS (Boroş et al., 2012). Toate alternativele generate de algoritmi au fost adaugate in lexicon, folosind HVite sa decida care este cea mai probabila transcriere fonetica.
2. Generarea unei transcrieri fonetice initiale (folosind HLEd) utilizand prima pronuntie disponibila din dictionar pentru fiecare cuvant din corpus.

3. Scanarea corpusului transcris fonetic din pasul anterior, generand date pentru modelul HMM astfel: o stare initiala (tri-stare), stanga-dreapta (model HMM monofon) pentru toti fonemii din baza de date (fara a include pauzele scurte – short pause / sp) folosind HERest, si reestimat modelul initial de 4 ori. Datele HERest sunt: '250.0 150.0 1000.0';
4. Adaugarea pauzei scurte (sp) in modelul HMM (initial copiat din modelul de liniste ‚sil' aflat de obicei la inceputul si finalul propozitiilor), si reestimarea modelului HMM nou creat de 4 ori. Se folosesc aceeasi parametrii HERest.
5. Re-generarea transcrierii fonetice (inclusiv generarea pauzelor scurte) din corpusul de voce (folosind HVite) folosind cel mai bun model HMM obtinut in pasul anterior – astfel se obtin cele mai bune pronuntari care corespund datelor acustice (in cazul in care cuvantul are mai multe forme de pronuntare)
6. In final, se reestimeaza (tot de 4 ori) modelul HMM monofon, incluzand pauzele scurte folosind noua transcriere a corpusului, generandu-se alinierile dorite, incluzand timpul start-stop pentru fiecare fonem.

4.5 A2.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii (partea a II-a) si A2.5. Implementarea unui controller de voce (partea I)

Din perspectiva automatizării şi integrării tehnologiilor ASR, TTS şi KNX, arhitectura centrală a proiectului ANVSIB este compusă din două module principale: (a) un proxy de voce (activitatea 2.5) şi (b) controllerul de voce. Deoarece aceste două sisteme sunt strâns legată unul de altul, în continuare vom prezenta cumulat procesul de implementare şi rezultatele obținute. Pentru o mai bună înțelegere a sistemului (Figura 3) vom explica în cele ce urmează modul de funcționare şi cerințele de sistem exemplificând acolo unde este cazul.

Unul din rolurile proxy-ului de voce este acela de a culege mai multe fluxuri de date (voce) şi de a gestiona transferul către controller în vedere convertirii fluxului audio în text. De asemenea, proxy-ul de voce este responsabil pentru prelurarea rezultatelor furnizate de controler şi pentru a fi redate pe terminalele conectate la proxy.

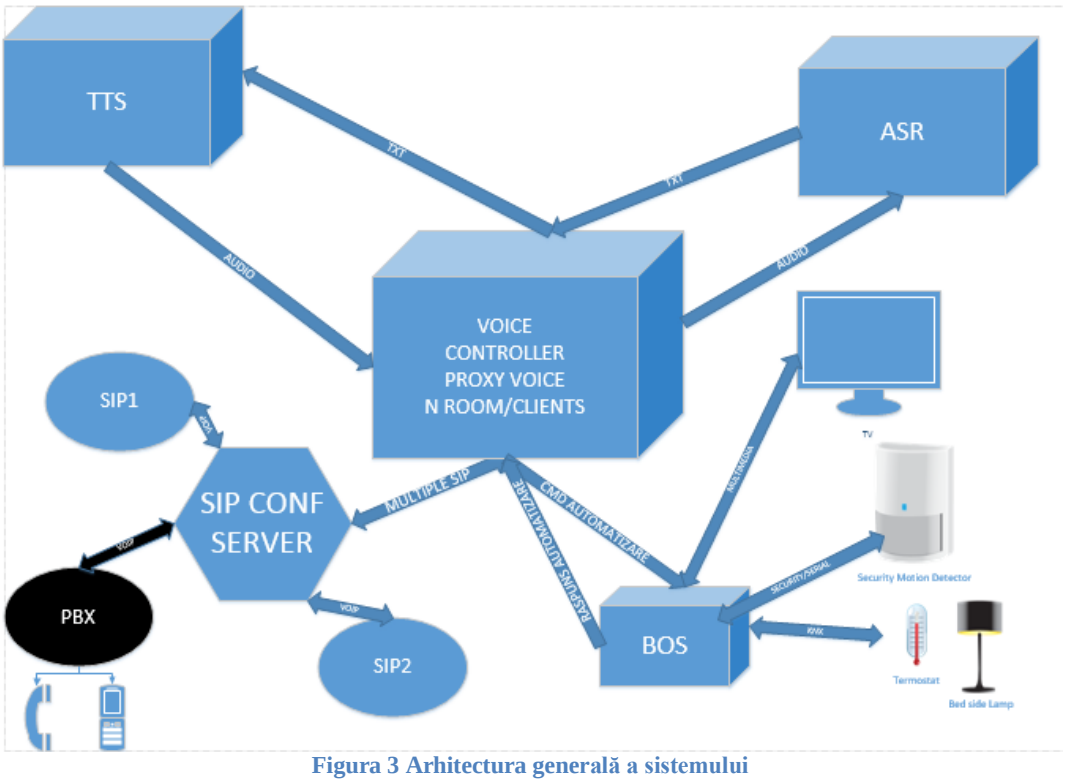

Figura 3 Arhitectura generală a sistemului

Controlerul de voce este destinat citirii de date din proxy, a comunicării cu sistemele TTS şi ASR şi a gestionării sistemului de automatizare. Aşa cum s-a menționat în raportul anterior, am ales metode şi tehnologii standard în implementarea prototipului ANVSIB: semnalizare SIP pentru terminale, transfer voce

prin RTP între terminale şi proxy (standard pentru telefonie IP), MRCP pentru comunicarea între controller şi sistemele TTS şi ASR. Pentru claritate, în continuare vom prezenta un flux de date complet:

a. Proxy-ul de voce asigura comunicarea bi-direcțională cu terminalele audio, folosind protocolul de semnalizare SIP. El transmite datele recepționate către controler folosind fişiere FIFO (specifice sistemelor de operare Unix şi Linux). Comunicarea FIFO (pipe) a fost aleasă pentru a micşora latența între cele două module.

b. Controlerul de voce primeşte date prin FIFO de la proxy şi retransmite aceste date către sistemul ASR folosind MRCP. El primeşte ca răspuns impoteza de decodare a sistemului ASR.

c. Pentru fiecare ipoteză primită, sistemul încearcă să identifice un scenariu de automatizare care se protriveşte cel mai bine cu textul primit. Odată identificat un scenariu el poate să ceară mai multe infomații sau să lanseze un script de automatizare care să comunice prin KNX cu dispozitivele din casă şi să returneze rezultat. Rezultatul este întotdeauna de tip text.

d. Rezultatul de tip text primit de la script este direcționat tot prin MRCP către sistemul TTS care în schimb returnează un flux audio ce reprezintă textul sintetizat.

e. Fluxul sintetizat este apoi redirecționat prin FIFO către proxy-ul de voce, care la rândul lui îl transmite mai departe către terminal, unde aceasta este auzit de utilizator.

În cele ce urmează vom prezenta modul în care fiecare subsistem este implementat, punând accent pe metode şi tehnicile folosite şi introducând librăriile externe utilizate. Astfel vom pune accent pe: (a) comunicarea SIP dintre proxy şi terminale; (b) comunicarea MRCP dintre controller, TTS şi ASR; (c) decodarea ipotezelor şi lansarea automatizărilor;

(a) Comunicarea SIP dintre proxy şi terminale

Aşa cum am menționat, proxy-ul de voce foloseşte protocolul SIP pentru semnalizare şi RTP pentru gestionearea fluxurilor de voce. El a fost implementat în JAVA şi utilizează librăria peers[1] pentru comunicație. Din punct de vedere al cerințelor de proiect, este necesară colectarea şi gestionarea mai multor fluxuri de voce. Pentru a menține o arhitectură clară şi a separa cât mult funcționalitățile de sistem, am ales implementarea proxy-ului de voce sub forma unui utilitar în consolă care, pe baza unui fişier de configurare, să gestioneze comunicația cu un singur flux audio. La pornire acest utilitar citeşte un fişier de configurare (vezi mai jos) din care îşi colectează datele de autentificare:

```
LOCAL_IP=xxx.xxx.xxx.xxx
GLOBAL_IP=xxx.xxx.xxx.xxx
USERNAME=1234
PASSWORD=**********
DOMAIN=xxx.xxx.xxx.xxx
```

Completarea informațiilor din configurare se face relativ uşor: LOCAL_IP – ip-ul local al clientului; GLOBAL_IP este IP-ul vizibil de către centrala SIP, el fiind diferit de LOCAL_IP în cazul în are se face traversarea unui NAT; USERNAME şi PASSWORD se referă la datele de autentificare cu centrala telefonică; DOMAIN reprezintă adresa IP sau numele de domeniu pentru centrala telefonică cu care se face înregistrarea.

La pornire se poate alege modul în care proxy-ul să acționeze: se stabileşte fişierul de configurație folosit, se aleg fişierele FIFO ce trebuie utilizate şi se stabilieşte dacă proxy-ul trebuie să inițieze el însuşi un apel către un terminal sau dacă trebuie să aştepte pentru conexiuni din afară. Pentru gesionarea mai multor terminale se folosesc mai multe instanțe ale aplicației.

(b) comunicarea MRCP dintre controller, TTS şi ASR

Aşa cum am menționat controlerul preia informații de la proxy-ul de voce şi le redirecționeaza către ASR şi TTS folosind MRCP. Implementarea este de asemenea realizată în JAVA şi pentru MRCP am folosit librăria UniMRCP2. La pornire aplicația citeşte un fişier de configurare de unde primeşte proxy-urile cu care trebuie comunice precum şi adresele sistemelor TTS şi ASR:

```
ASR_ADDR=xxx.xxx.xxx.xxx
TTS_ADDR=xxx.xxx.xxx.xxx
```

1 http://peers.sourceforge.net/

2 http://www.unimrcp.org/

```
STREAM_COUNT=n
STREAM 1 <fifo_in><fifo_out>
STREAM 2 <fifo_in><fifo_out>
.
.
.
STREAM n <fifo_in><fifo_out>
```

În mod standard, primele două câmpuri se completează cu adresele IP ale host-urilor ce ruleaza sistemle ASR, respectiv TTS. În continuare, controlerul de voce citeşte numărul de fluxuri audio pe care trebuie să le gestioneze (STREAM_COUNT), urmând ca ulterior, pe următoarele linii să fie regăsite denumirile fişierelor FIFO pentru fiecare flux în parte.

(c) decodarea ipotezelor şi lansarea automatizărilor

În momentul de faţă decodarea ipotezelor primite de la sistemul ASR se face utilizând un sistem bazat pe reguli. Acest sistem a fost detaşat de restul arhitecturii şi, pentru o testare rapidă a fost implementat în PHP. Alegerea unui limbaj interpretat pentru dezvoltare este motivată de faptul că se pot face modificări în timp real asupra codului, fără a fi necesară o recompilare şi repunere în lucru a sistemului de decizie. Acest lucru ne va permite în viitor efectuarea rapidă a testelor. Odată idenficat un scenariu, decodorul scris în PHP lansează în execuţie un script aferent, ce are ca scop comunicarea prin standardul KNX şi lansarea în execuţie a operaţiilor de control sau interogare a dispozitivelor din casă. Script-urile de automatizare sunt executabile de tip consolă ce returnează prin ieşirea standard text un mesaj. Acest mesaj este preluat de controler, sintetizat folosind sistemul TTS şi retrimis către proxy-ul de la care a plecat fluxul pentru a ajunge în final la terminalul de la care s-a iniţiat cererea utilizatorului.

4.6 A2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de cameră) (partea a II-a)

Baza de date ce a fost achiziţionată în cadrul acestei activităţi are patru părţi. Cele patru părţi au fost realizate în vederea realizării mai multor aplicaţii de procesare a vorbirii, fiecare cu necesităţile ei. În cele ce urmează vom trata pe rând fiecare parte a bazei de date.

Baza de date de comenzi pentru automatizarea unei case inteligente. S-a dorit crearea unei baze de date care să cuprindă o serie de comenzi ce pot fi date de utilizator casei sale inteligente. Un anumit grad de variabilitate a fost luat în considerare, gestionând lista comenzi în aşa fel încât să acopere diferite moduri de a exprima aceeaşi comandă şi fiecare comandă a fost rostită un anumit număr de ori, în funcţie de frecvenţa cu care este de aşteptat să fie rostită într-un scenariu real. Prin urmare, au fost înregistrate comenzi rostite de fiecare vorbitor, în fiecare cameră a casei, la diferite distanţe faţă de microfoane. Acest lucru ne va oferi informaţii despre multe aspecte care influenţează rezultatele recunoasterii vorbirii la distanţă (RVD), şi anume: modul în care calitatea înregistrării, distanţa dintre vorbitor şi microfoane, orientarea vorbitorului faţă de microfon, reverberaţia camerei, ecoul camerei, zgomotul ambiental interior si exterior.

Baza de date de tuse pentru detecţia de crize de tuse. Această bază de date poate fi utilizată pentru detectarea situaţiilor în care cineva suferă de o tuse acută şi sistemul ar trebui să detecteze automat acest lucru şi trimite un mesaj de alarmă către cineva care poate veni pentru a ajuta. Pentru a evita confuzia cu tusea normală care poate apărea din diverse motive nesemnificative, vorbitorii au fost rugati să tuşească scurt, lung şi să-şi "dreagă vocea". Astfel de sisteme de detecţie pot fi utile pentru persoanele care suferă de afecţiuni pulmonare sau persoane care se îneacă şi se pot sufoca şi au nevoie de asistenţă urgentă.

Baza de date de vorbire spontană, la distanţă. Această bază de date poate fi utilizată pentru evaluarea sistemelor de recunoaştere (transcriere) de vorbire la distanţă. Această bază de date va fi combinată cu baza de date de comenzi pentru a stabili cum se comportă sistemul când utilizatorul rosteşte o comandă în mijlocul unei conversaţii. De asemenea, folosind acest corpus se pot dezvolta aplicaţii de identificare şi verificare de vorbitor în vederea oferirii de securitate sporită caselor inteligente.

Baza de date multi-vorbitor, multi-lingvă, multi-cameră. Această bază de date poate fi utilizată pentru stabilirea numărului de persoane dintr-o cameră. Acest scenariu, denumit în limbaj tehnic "scenariu de tip cocktail party", poate fi adresat prin tehnici de separare a surselor acustice. Pe lângă condiţiile deja adverse pe care situaţiile de cocktail-party le presupun, camere multiple prin care vorbitorii se deplasează, microfoanele multiple care înregistrează în acelaşi timp şi limbile multiple în care se vorbeşte (engleză,

franceză şi română) au fost adăugate la lista de variabile. Aplicația de identificare a numărului de persoane dintr-o cameră nu trebuie să ia în considerare limba în care se vorbeşte, dar acest lucru poate fi utilizat pentru alte experimente.

Toate bazele de date de vorbire conțin vorbitori de sex masculin şi feminin, astfel încât variabilitatea de gen poate fi, de asemenea, studiată. După cum se poate observa din ceea ce a fost deja prezentat, cele patru seturi de date care au fost înregistrate pot fi ținta a numeroase aplicații în domeniul RVD, în funcție de contextul şi de obiectivul cercetării. Chiar dacă principalele aplicatii din acest proiect au fost clar definite pentru baza de date care a fost creată, există şi alte tipuri de proiecte şi aplicații care pot beneficia de ea. De asemenea, este demn de menționat faptul că nu sunt multe medii disponibile şi deja funcționale ca cel în care au fost făcute înregistrările, şi anume casa inteligentă DOMUS deținută de Laboratorul de Informatică din Grenoble.

4.6.1 Descrierea mediului de înregistrări (casa inteligentă DOMUS)

Casa inteligentă DOMUS, deținută de Laboratorul de Informatică din Grenoble, este complet echipată cu senzori şi actuatori, oferind un mediu cu inteligență ambiantă şi cuprinde o bucătarie, un dormitor, o baie şi un birou, aşa cum este prezentat în Figura 4. Toate camerele dispun de mobilierul necesar, aparate de uz casnic şi asigură un anumit grad de confort, astfel încât apartamentul să fie complet utilizabil. Casa inteligentă oferă o serie de sisteme, cum ar fi jaluzele, perdele, televizor, încuietori pentru uşi etc., care pot fi controlate printr-o interfață online, prin urmare comenzile pe care utilizatorul le dă casei fac referire la aceste utilități. Pereții şi tavanul sunt realizate din material izolator audio pentru a menține zgomotul exterior la un nivel minim. Cu toate acestea, ferestrele din sticlă nu împiedică zgomotul ambiental exterior să interfereze cu înregistrările. De asemenea, casa prezintă şase camere video, câte două în fiecare încăpere, cu excepția băii, care din motive de intimitate a rămas neechipată cu o astfel de cameră. Scopul camerelor video este monitorizarea mişcării şi poziției oamenilor de-a lungul experimentului. Aceste materiale filmate sunt folosite pentru a verifica dacă adnotarea bazei de date a fost corectă din punct de vedere al camerei în care vorbitorii s-au aflat la momentul înregistrării şi sunt disponibile în baza de date pentru orice posibilă utilizare ulterioară referitoare la problemele de adnotare.

Pentru realizarea înregistrărilor audio, apartamentul este echipat cu şapte microfoane SENNHEISER ME2 (câte 2 în fiecare cameră, cu excepția băii, unde este prezent numai unul singur), fixate în tavan. Din cele şapte microfoane, patru oferă înregistrări de o calitate superioară, aflându-se câte unul in fiecare cameră. Toate acestea sunt conectate la o placă de achiziție de date multicanal, de tip National Instrument PCI-6220E, cu ajutorul căreia se poate înregistra semnalul audio în timp real, simultan pentru toate microfoanele [UPB1]. Software-ul care a fost folosit pentru a efectua înregistrări pe toate canalele în acelaşi timp a fost StreamHIS [UPB2]. Instrumentul folosit pentru transcrierea înregistrărilor audio a fost Transcriber (http://trans.sourceforge.net). În fiecare sesiune, întregul semnal audio a fost capturat de către toate cele şapte microfoane în acelaşi timp şi toate aceste înregistrări sunt disponibile în baza de date.

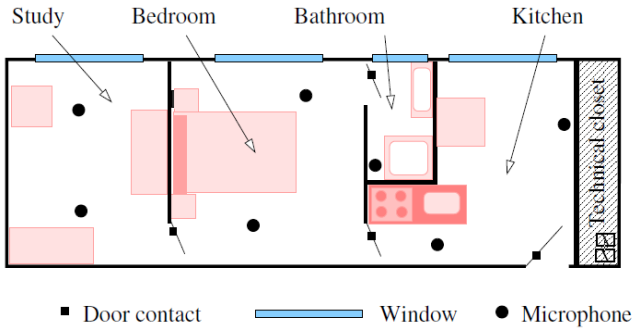

Figura 4 Descrierea casei inteligente DOMUS [UPB1]

4.6.2 Metodologia de achiziție e bazei de date

Primul pas în procesul de achiziție a bazelor de date a fost de a stabili obiectivele privitoare la ceea ce se doreşte a se obține în cele din urmă. Prin urmare, fiecare dintre cele patru baze de date au fost înregistrate pornind de la nişte scenarii clare, aşa cum vor fi descrise în secțiunea următoare.

Urmatorul pas a fost de a aduna vorbitori de limba română pentru crearea înregistrărilor contactând mai mulți studenți români din Grenoble. Ei au fost îndeajuns de binevoitori pentru a ajuta la realizarea înregistrărilor şi chiar au contribuit la strangerea mai multor persoane in acest scop. Gama de vârsta a

acestora variază între 22 şi 41 de ani, aşa cum se poate vedea în Tabelul 1. Unii vorbitori nu au putut înregistra partea de bază de date de vorbire spontană sau pe cea multi-vorbitor, multi-lingvă, multi-cameră din cauza lipsei de sincronizare între programul lor şi intervalele de timp când mediul de înregistrări a fost disponibil.

Etapa finală a reprezentat întâlnirea cu voluntarii şi crearea înregistrărilor în sine, care a început prin explicarea termenilor şi condiţiilor în care aceste experimente vor avea loc. Ei au fost rugaţi să semneze un formular de consimţământ (document in extenso nr. 5) privind atât înregistrările audio cu ajutorul microfoanelor, cat şi înregistrările video cu camerele de luat vederi. Ei au primit o copie semnată a acestor acorduri, iar o altă copie a rămas la echipa de organizare.

Înainte de a începe fiecare experiment, voluntarii au dat un tur complet al casei şi le-au fost prezentate toate încăperile şi sistemele automatizate (document in extenso nr. 6). Apoi, li s-a cerut să citească cu voce tare acordul de consimţământ (document in extenso nr. 5), în interiorul bucătăriei, pentru adaptarea modelului acustic. Acest text scurt conţine toate fonemele limbii române. În timpul achiziţiei primelor 2 baze de date, aceştia au fost singuri în apartament şi toate uşile au fost deschise.

Tabelul 1 Detalii demografice ale vorbitorilor

Vorbitori primari				
Id vorbitor	***Sex***	***Vârstă***	***Înălţime***	***Greutate***
S1	M	24	1.87	90
S2	F	23	1.64	54
S3	M	24	1.87	90
S4	F	23	1.69	67
S5	F	23	1.58	53
S6	M	21	1.77	66
S7	F	23	1.70	57
S8	F	41	1.80	84
S9	F	30	1.59	54
S10	F	24	1.58	57
S11	F	21	1.60	56
Vorbitori auxiliari				
S12	M	35	N/A	N/A
S13	M	51	N/A	N/A
S14	F	42	N/A	N/A
S15	M	24	N/A	N/A

4.6.2.1 Baza de date de comenzi

Pentru baza de date de comenzi a fost întocmită o listă care cuprinde toate utilităţile posibile ce pot fi operate de către sistemul central de control disponibil în casa in care acest sistem de RVD va fi folosit. În scopul de a ajuta oamenii să se obişnuiască cu casa, în cadrul comenzilor a fost introdus un cuvânt cheie ce reprezintă numele casei inteligente, şi anume Casandra. Acesta este primul cuvânt din fiecare comandă, aşa cum se vede în Figura 5, un fragment din gramatica cu stări finite (FSG), care stă la baza listei de comenzi. Pornind de la această gramatică, a fost dezvoltată o listă de comenzi posibile care pot fi rostite de un vorbitor ce doreşte să controleze sistemele de automatizare. Aceste comenzi au fost împărţite în 3 categorii, în funcţie de frecvenţa cu care se aşteaptă a fi rostite într-un scenariu de caz real.

Comenzile cele mai probabile (de exemplu, aprinderea/stingerea luminii) au fost rostite de trei vorbitori diferiţi, comenzile mai puţin probabile (de exemplu, deschiderea/închiderea jaluzelelor) au fost rostite de către doi vorbitori diferiţi, iar cele mai puţin probabile (de exemplu, pornirea/oprirea simulatorului de prezenţă) doar de către un vorbitor. Fiecarui voluntar i s-a dat un set de aproximativ 45 de comenzi intercalate cu unele fraze care nu sunt legate de inteligenţa ambiantă, într-un raport de 7:1, pentru a putea evalua unele fraze care nu sunt comenzi înregistrate în acelaşi mediu, în condiţii similare. Ei au fost rugaţi sa repete acest set în fiecare cameră, cu excepţia băii, unde setul a fost redus la un sfert, având în vedere faptul că acolo este mai puţin probabil să existe aceeaşi cantitate de activitate vocală. Participanţilor nu li s-a dat nicio instrucţiune cu privire la care ar trebui să fie orientarea lor spre microfon, nici nu au fost restricţionati în ceea ce priveşte amplasarea lor în cameră în momentul vorbirii. De asemenea, nu le-a fost limitată posibilitatea de a sta jos la masă, pe pat sau pe canapea şi nu le-a fost comunicată poziţia microfoanelor din casă până la sfârşitul experimentului, cu excepţia cazului în care au cerut în mod special acest lucru. Li s-a

spus numai să vorbească cu o voce normală, la fel cum ar vorbi cu o altă persoană, cu o scurtă pauză (1-2 secunde) între comenzi şi să repete comenzile dacă simt că este necesar.

Toți vorbitorii au urmat același scenariu (documentele in extenso 2, 3, 4), ce le-a fost oferit pe o foaie de hârtie, cu privire la comenzile pe care ar trebui să le rostească şi camera în care ar trebui să se afle în fiecare moment. Pentru fiecare vorbitor sunt disponibile şapte înregistrări complete, corespunzătoare fiecărui microfon din casa inteligentă. Pentru adnotarea vorbirii, cele mai bune înregistrări au fost selectate în funcție de poziția în care vorbitorul s-a aflat atunci când a rostit comenzile. Prin urmare, pentru o mai bună înțelegere a ceea ce a fost rostit, semnalul audio înregistrat de microfonul din bucătărie în timp ce vorbitorul s-a aflat în bucătărie a fost unit cu semnalul audio de la microfonul din dormitor, în timp ce vorbitorul s-a aflat în dormitor, in mod similar procedându-se cu semnalul audio înregistrat de microfonul din baie, precum şi cel din camera de studiu. Pentru a evita problemele de sincronizare, nicio secvență audio din secțiunea de comenzi nu a fost tăiată. Starea finală a bazei de date este detaliată în Tabelul 2.

```
<electric> = <lumini> | <culoare_lumina> | <dimmer> | <alimentare> |
             <panouri_solare>;

<lumini> = (aprinde|stinge|oprește|pornește|închide|deschide)
             (lumina|[toate] luminile) [[de] la parter|din casă];
<culoare_lumina> = (schimbă|aprinde) (culoarea luminii|lumina)
             (roșie|în roșu|[în] verde|[în] albastru|normală|la
             normal);
<dimmer> = (mărește|crește|micșorează|scade) luminozitatea la
             <numar_zeci> la sută;
<alimentare> = (pornește|reia|oprește) (alimentarea|încărcarea
             bateriei);
<panouri_solare> = care este starea bateriei|este încărcată bateria;
//-------------------------------------------------------------------
public <comandaCasa> = casandra (<exterior> | <securitate> |
             <multimedia> | <hvac> | <electric>);
```

Figura 5 Un fragment din gramatica cu stări finite (FSG) care stă la baza listei de comenzi

Tabelul 2 Detalii privind bazele de date de vorbire înregistrate

Numele bazei de date	# vorbitori	Durata înregistrărilor	Observații
Comenzi	11	00:58:05	în total 1764 de comenzi
Tuse	11	00:06:10	în total 240 de secvențe audio conținând tuse („dregerea glasului”, tuse scurtă şi tuse lungă)
Vorbire spontană la distanță	8	01:53:50	dependență mare de dorința vorbitorilor de a da răspunsuri ample; doar vorbirea provenind de la doi vorbitori a fost anotată
Multi-vorbitor, multi-linvă, multi-cameră	12	00:53:46	dependență mare de dorința vorbitorilor de a da răspunsuri ample

4.6.2.2 Baza de date de tuse

Pentru baza de date de tuse a fost oarecum dificil să se diferențieze o tuse uşoară de ceea ce poate fi considerată o tuse critică, gravă, care poate să amenințe sănătatea cuiva. De aceea s-au realizat trei sesiuni separate de tuse, fiecare dintre ele constând într-o dregere a vocii, o tuse scurtă şi una lungă. După ce s-a observat că este obositor să se efectueze aceste sesiuni în mod repetat, fără a lua o pauză, au fost impuse unele restricții în acest sens, intercalând sesiunile cu înregistrări de comenzi, pentru ca cei care vorbesc să nu îşi piardă vocea sau să ajungă la epuizare. Aceste restricții au fost urmate în cele mai multe cazuri, de unele greşeli minore, şi anume cazuri în care vorbitorul a uitat să efectueze o tuse lungă sau a tuşit scurt atunci când ar fi trebuit să efectueze o tuse lungă. Cu toate acestea, excepțiile respective au fost tratate şi adnotate corespunzător. Un aspect care a fost remarcat şi poate fi util în cercetările ulterioare este că tusea lungă este foarte asemănătoare cu o tuse scurtă repetată. Acest lucru depinde de faptul că limita maximă de aer pe care plămânii umani o pot expira este de ajuns doar pentru o tuse scurtă, de aceea pentru o tuse lungă este repetat acest proces de mai multe ori. Din păcate, nicio sesiune de tuse naturală nu a putut fi înregistrată deoarece toți voluntarii au fost sănătosi, dar ceea ce au reuşit să simuleze este aproape de scenariile reale. În cele din urmă, fiecare vorbitor a înregistrat 2 sesiuni de tuse în fiecare cameră, cu excepția băii în care a avut loc una singură, fiind stabilite perioade mai lungi de timp între ele, astfel încât vorbitorii să îşi poată odihni vocea.

Pentru adnotare, a fost aplicată aceeaşi tehnică ca şi în cazul precedent. Detalii despre ce a fost înregistrat pot fi vizualizate în Tabelul 2.

4.6.2.3 Baza de date de vorbire spontană la distanță

Pentru baza de date de vorbire spontană a fost alcătuită o listă de întrebări (document in extenso nr. 7) care au fost adresate fiecărui vorbitor care a participat la acest scenariu, cu scopul de a furniza cât mai multă vorbire liberă. Lista de întrebări a fost concepută în aşa fel încât să fie posibil ca fiecare să ofere răspunsuri lungi şi complexe. Întrebările nu au fost intruzive sau prea personale astfel incât să nu îi facă pe vorbitori să se simtă inconfortabil. La început, a existat o încercare de a adresa întrebările prin intermediul unui laptop, dar acest lucru ar fi însemnat un consum de timp şi vorbitorii nu ar fi oferit răspunsuri suficient de lungi, iar comunicarea ar fi fost oarecum artificială. Prin urmare, unul dintre organizatori a stat cu voluntarul în cauză şi a adresat aceste întrebări, păstrând însă intervențiile sale la un nivel minim. Astfel el a putut să adreseze întrebări suplimentare, bazate pe răspunsurile primite de la participanți, dialogul fiind mai bogat si mai lung. Detaliile cu privire la această bază de date pot fi găsite în Tabelul 2.

4.6.2.4 Baza de date de vorbire multi-vorbitor, multi-lingvă, multi-cameră

Pentru baza de date multi-speaker, multi-lingvă, multi-cameră, au fost proiectate mai multe scenarii (documentele in extenso 2, 3, 4) astfel încât acestea să garanteze prezența unui număr diferit de persoane în fiecare cameră, unele tranziții de la o cameră la alta, cu uşile atât deschise cât şi închise, înregistrându-se discursuri în mai multe limbi, de vorbitori de ambele sexe. Aceste înregistrări au fost efectuate în interiorul unui apartament, cu ajutorul a 4 persoane, împărțite în 2 grupuri. Pentru a se asigura că totul decurge fără probleme unul dintre organizatori a fost liderul experimentului, fiind, de asemenea, parte a unuia dintre cele două grupuri. Scenariul a decurs, după cum urmează: toate cele 4 persoane au pornit de la bucătărie şi s-a discutat timp de 1-2 minute, cu uşa între bucătărie şi dormitor închisă. Apoi, liderul a luat o persoană cu el şi s-au dus în birou, închizând uşa la bucătărie în spatele lor. Din motive tehnice, uşa dintre dormitor şi birou nu a putut fi închisă, astfel rămânând deschisă în timpul întregului experiment. O dată ce grupurile s-au separat în bucătărie şi, respectiv, birou, au avut loc convorbiri de încă un minut în ambele cazuri. După aceea, liderul a rugat pe cineva de la bucătărie să deschidă uşa spre dormitor şi conversațiile au continuat în fiecare cameră, ca şi înainte, timp de un minut. Acest lucru a fost făcut pentru a avea loc şi interferențe între camere cu scopul de a avea la dispoziție pentru cercetări viitoare cât mai multe tipuri posibile de vorbire. Apoi liderul a mers în dormitor şi simulat o convorbire la telefon cu un prieten, în timp ce grupul de la bucătărie a continuat conversația şi cel rămas în birou a păstrat liniştea. Apoi, liderul a început să cheme restul oamenilor în dormitor să se alăture unei conversații începând cu persoana care a rămas singură în birou, urmată de persoanele din bucătărie, lăsând, de asemenea, uşa la bucătărie deschisă. După ce una dintre persoane a intrat în dormitor, grupul a susținut o scurtă conversație de aproximativ 1 minut, până când liderul a chemat următoarea persoană. Unele subiecte de discuție (document in extenso nr. 1) au fost alese înaintea experimentului, pentru a evita situațiile în care oamenii nu ar avea despre ce să vorbească, dat fiind faptul că s-au cunoscut abia cu câteva minute înainte de experiment. Aceste subiecte au fost tipărite şi oferite voluntarilor. Figura 6 prezintă un moment din acest experiment, în care trei persoane sunt înregistrate de camerele video în interiorul dormitorului.

Faptul că era nevoie ca toate persoanele să vorbească într-o perioadă de timp scurtă a fost, de asemenea, o provocare, deoarece acest lucru depinde de personalitatea acestora, astfel liderul fiind nevoit să pună întrebări individual fiecăruia în scopul de a aduna vorbire de la fiecare participant în parte. O sesiune de înregistrare multi-speaker, multi-lingvă, multi-cameră, a durat aproximativ 10 de minute şi nu au existat constrângeri în ceea ce priveşte poziția, intensitatea sau subiectul conversației. Scopul acestei baze de date a fost de a achiziționa semnalul de vorbire de la mai multe persoane plasate în camere diferite. Folosirea mai multor limbi utilizate în timpul înregistrărilor nu ar trebui să aibă un impact puternic asupra rezultatelor. Tabelul 2 oferă informații despre ceea ce a fost înregistrat în toate cele patru baze de date.

Figura 6 Exemplu de înregistrări preluate de camerele video din casa inteligentă

4.7 A2.7. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)

Reducerea zgomotului şi îmbunătățirea semnalului acustic sunt două dintre cele mai active domenii de cercetare, în recunoaşterea automată a vorbirii (RAV). Aceasta este, de fapt, zona în care sistemele de RAV sunt complet depăşite de capacitatea umană de a recunoaşte vorbirea. În ultimul deceniu, robustețe la zgomot a sistemelor RAV a fost abordată în diverse feluri: prin îmbunătățirea semnalului acustic, cu ajutorul parametrilor vocali robuşti la zgomot, prin modele acustice adaptate la zgomot, etc. Dintre toate aceste tehnici, utilizarea parametrilor Power Normalized Cepstral Coefficients (PNCCs), pare să aducă cele mai importante câştiguri în precizie. Un atribut cheie al acestor parametri este că ei folosesc a) analiza de putere pe termen mediu, în care parametrii de mediu sunt estimați pe o durată de timp mai mare decât cea utilizată în mod frecvent pentru procesarea vorbirii, şi b) frecvența de uniformizare. Alte atribute distinctive majore de prelucrare PNCC includ:

- utilizarea nonliniarității de tip power-law care înlocuieşte nonliniaritatea de tip logaritmică utilizată tradițional în calculul coeficienților Mel Cepstrali (Mel Frequency Cepstral Coefficients MFCC);
- un algoritm de reducere a zgomotului bazat pe o filtrare asimetrică ce elimină zgomotul de fond;
- un modul de mascare temporală.

Având în vedere acest context, precum şi rezultatele bune raportate pentru alte limbi, am decis să implementăm parametrii vocali PNCC în sistemul nostru de RAV. Am contribuit la implemnetarea acestor parametrii robuşti la zgomot în CMU Sphinx, un toolkit open-source de recunoaştere a vorbirii, care este nucleul sistemului nostru de RAV. Apoi, am experimentat cu aceşti noi parametri ca înlocuitori pentru parametrii MFCC tradiționali şi am obținut rezultate mai bune pentru toate tipurile de vorbire zgomotoasă pe care am încercat să o transcriem.

Efectul utilizării parametrilor vocali robuşti în locul parametrilor vocali tradiționali (MFCC) a fost evaluat pe:

- RSC-eval (read-speech corpus), un corpus de vorbire continuă, citită, în limba română, cuprinzând înregistrări curate (fără zgomot), creat de grupul de cercetare SpeeD înainte de începerea acestui proiect;
- SSC-eval (spontaneous-speech corpus), un corpus ce conține vorbire spontană înregistrată şi în condiții de linişte şi în condiții de zgomot (zgomot sau muzică de fundal, vorbire la telefon, etc.), de asemenea creat de grupul de cercetare SpeeD înainte de începerea acestui proiect.

Evaluarea a fost realizată cu ajutorul unui model acustic antrenat cu parametrii vocali PNCC (AM001) şi un model de acustic antrenat cu parametrii vocali MFCC (AM002).

Pentru a evalua robustețea sistemul de RAV în diferite condiții de zgomot, mai multe tipuri de zgomot (stradal, murmurat şi de metrou) au fost adăugate digital, la diferite raporturi semnal-zgomot (SNR – signal to noise ratio), peste vorbirea curată din RSC-eval. Figura 7 ilustrează rezultatele obținute cu modelul acustic bazat pe MFCC (AM002) şi modelul acustic bazat pe PNCC (AM001) pe corpusurile RSC-eval cu zgomot

adăugat digital. Figura arată că, indiferent de tipul de zgomot şi indiferent de SNR, modelul acustic bazat pe PNCC este mai bun decât modelul acustic bazat pe MFCC. Chiar şi pentru vorbirea curată modelul acustic bazat pe PNCC obţine un WER (word error rate) ceva mai bun. Reducerea relativă WER este, în medie, 10% pentru un SNR de 20dB, 16% pentru un SNR de 15dB, 19% pentru un SNR de 10dB, şi 13% pentru un SNR de 5dB.

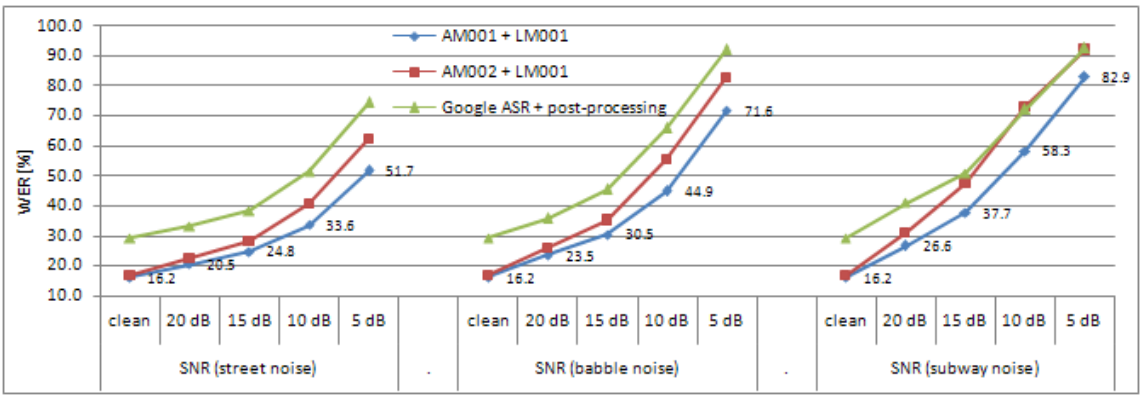

Figura 7 Îmbunătăţiri RAV aduse de parametrii vocali PNCC + comparaţie cu sistemul RAV al Google

Condiţiile de vorbire din corpusul SSC-eval sunt adnotate la nivel de propoziţie şi acest lucru ne-a permis să împărţim corpusului în două părţi: vorbire curată şi vorbire zgomotoasă. In Tabelul 3, se evaluează modelul acustic bazat pe MFCC (AM002) şi modelul acustic bazat pe PNCC (AM001) atât pe corpusul SSC-eval ca un întreg, cât şi separat pe partea sa curată şi pe partea sa zgomotoasă. Acest experiment are ca scop să evalueze robusteţea la zgomot a sistemului RAV în condiţii de zgomot din lumea reală. Rezultatele arată că reducerea relativă de WER pe vorbire zgomotoasă (SSC-eval zgomotos) este mai mică (4%) decât în experimentul anterior. Un alt fapt interesant este că pentru vorbirea curată (SSC-eval curat) modelul acustic robust la zgomot (AM001) obţine un WER puţin mai slab (o eroare mai mare) decât modelul acustic bazat pe MFCC. Cu toate acestea, modelul acustic bazat pe PNCC este de per ansamblu mai bun decât modelul acustic bazat pe MFCC (SSC-eval întreg).

Tabelul 3 Îmbunătăţiri RAV aduse de parametrii vocali PNCC

		WER [%]			
Modelul acustic	**Parametri vocali**	**RSC-eval**	**SSC-eval (întreg)**	**SSC-eval (curat)**	**SSC-eval (zgomot)**
AM002	MFCC	16.7	39.4	31.4	51.1
AM001	PNCC	16.2	38.9	31.7	49.1

4.8 Sumar al realizărilor / rezultatelor

Nr.	Denumirea rezultatului	Denumire activitate	Procent de realizare
1.	*D2.1 - Corpus de vorbire anotat si aliniat pentru limba romana*	*A2.3 Segmentarea si adnotarea corpusului de vorbire (partea I)*	*100%*
2.	*D2.2 - Model functional pentru proxy-ul de voce*	*Activitatea 2.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii (partea a II-a)*	*100%*
3.	*D2.3 - Model functional pentru controlerul de voce*	*Activitatea 2.5. Implementarea unui controller de voce (partea I)*	*100%*
4.	*D2.4 - Baza de date de vorbire inregistrata in conditii reale de camera*	*A2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de camera) (partea a II-a)*	*100%*
5.	*D2.5 - Model functional de recunoastere a vorbirii la distanta, robust la zgomot*	*A2.7. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea I)*	*100%*

4.9 Bibliografie

[1] Rehm, G., & Uszkoreit, H. (2012). META-NET White Paper Series: Europe's Languages in the Digital Age.

[2] Young, S., Woodland, P. C., & Byrne, W. J. (1993). HTK: Hidden Markov Model Toolkit V1. 5.

[3] Ratnaparkhi, A. (1996, May). A maximum entropy model for part-of-speech tagging. In Proceedings of the conference on empirical methods in natural language processing (Vol. 1, pp. 133-142).

[4] Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1), 39-71.

[5] Marques, N. C., & Lopes, G. P. (1996). A neural network approach to part-of-speech tagging. In Proceedings of the 2nd Meeting for Computational Processing of Spoken and Written Portuguese (pp. 21-22).

[6] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[7] Tufiş, D. (1999). Tiered tagging and combined language models classifiers. In Text, Speech and Dialogue (pp. 843-843). Springer Berlin/Heidelberg.

[8] Crammer, K., & Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. The Journal of Machine Learning Research, 3, 951-991.

[9] Boroş, T., Ştefănescu, D., & Ion, R. (2012). Bermuda, a data-driven tool for phonetic transcription of words. In Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop Programme (p. 35).

[10] Boros, T., Ion, R., & Tufis, D. (2013a). Large tagset labeling using Feed Forward Neural Networks. Case study on Romanian Language. In ACL (1) (pp. 692-700).

[11] Boros, T. (2013b). A unified lexical processing framework based on the Margin Infused Relaxed Algorithm. A case study on the Romanian Language. In RANLP (pp. 91-97).

[UPB1] Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, et al.. The Sweet-Home speech and multimodal corpus for home automation interaction. The 9th edition of the Language Resources and Evaluation Conference (LREC), May 2014, Reykjavik, Iceland. pp. 4499-4506.

[UPB2] Benjamin Lecouteux, Michel Vacher, and François Portet. Distant speech recognition for home automation: Preliminary experimental results in a smart home. In IEEE SPED 2011, pages 41–50, Brasov, Romania.

5 Documente in extenso

Nr.	Denumire anexă	Activitate asociata	Tip*
1.	*Conversation_Topics_Suggestions.docx*	*A2.6. Achizitionarea unei baze de date de vorbire in conditii reale (de camera) (partea a II-a)*	
2.	*Group_WO_Leader_Scenario.docx*		
3.	*Leader_Accompaniament_Scenario.docx*		
4.	*Leader_Scenario.docx*		
5.	*Consent_form.doc*		
6.	*Protocol.docx*		
7.	*Spontaneous_speech_discussion_topics.docx*		

6 Diseminarea rezultatelor proiectului

O parte din rezultatele proiectului au fost diseminate în două articole trimise la reviste indexate ISI. Unul dintre acestea a fost publicat în numărul din luna august 2015 al revistei „Revista Tecnica de la Faculdad de Ingineria Universidad del Zulia", iar celălalt va fi publicat în numărul din luna decembrie 2015 a revistei "Buletinul Ştiinţific UPB, Seria C". Referinţele complete ale articolelor se găsesc mai jos. Pentru referinţe actualizate se va consulta pagina web a proiectului de cercetare.

- Alexandru Caranica, Horia Cucu, Andi Buzo, Corneliu Burileanu, “Exploring Spoken Term Detection with a Robust Multi-Language Phone Recognition System,” in Revista Tecnica de la Faculdad de Ingineria Universidad del Zulia, vol. 38(2), pp. 97 - 104, August 2015.
- Valentin Andrei, Horia Cucu, Andi Buzo, Corneliu Burileanu, “Perception study inspired method for automatically detecting the number of competing speakers,” in University “Politehnica” of Bucharest Scientific Bulletin, Series C, accepted, in press.

În afara celor menționate mai sus proiectul a fost făcut cunoscut în forumurile în care laboratorul SpeeD are acces. Adresa paginii web a proiectului a rămas neschimbată: **http://speed.pub.ro/anvsib.**