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: *20.10.2016*

Etapa: *3/2016*

Activitatea / activităţile:

- *Activitatea 3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)*
- *Activitatea 3.2 Validarea manuala a corpusului de vorbire*
- *Activitatea 3.3. Implementarea unui controller de voce (partea a II-a)*
- *Activitatea 3.4. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea a II-a)*
- *Activitatea 3.5. Compararea si evaluarea performantelor tehnicilor implementate*
- *Activitatea 3.6. Diseminarea rezultatelor proiectului (partea I)*
- *Activitatea 3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti*
- *Activitatea 3.8. Dezvoltarea unor modele acustice multilingve si pentru limba engleza (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 treia fază a proiectului ANVSIB, obiectivele UPB au fost a) finalizarea modelului functional de recunoastere a vorbirii la distanta, robust la zgomot şi b) studiul aplicativ al algoritmilor state-of-the-art pentru detecția multilingvă a termenilor vorbiți.

În a treia fază a proiectului ANVSIB, obiectivele RACAI au fost segmentarea si adnotarea corpusului de vorbire si validarea lui manualapentru a produce un model acustic de calitate.

În a treia fază a proiectului ANVSIB, obiectivul IWAVE a fost finalizarea implementării pentru controllerul de voce (model funcțional).

2 Perioada de desfăşurare şi personalul implicat

Nr.	Denumire activitate	Nume si prenume	Perioada
1.	*Activitatea 3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş*	*oct 2015 – aprilie 2016*
2.	*Activitatea 3.2 Validarea manuala a corpusului de vorbire*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş*	*mai 2016 – august 2016*
3.	*Activitatea 3.3. Implementarea unui controller de voce (partea a II-a)*	*Badulescu Marian*	*noiembrie 2015 – septembrie 2016*
4.	*Activitatea 3.4. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea a II-a)*	*Cucu Horia; Buzo Andi; Burileanu Corneliu; Banica Alina; Minca Florentina*	*octombrie 2015 – decembrie 2015*
5.	*Activitatea 3.5. Compararea si evaluarea performantelor tehnicilor implementate*	*Cucu Horia; Buzo Andi; Burileanu Corneliu; Banica Alina; Minca Florentina*	*ianuarie 2016 – iunie 2016*
6.	*Activitatea 3.6. Diseminarea rezultatelor proiectului (partea I)*	*Cucu Horia; Buzo Andi; Burileanu Corneliu; Banica Alina; Badulescu Marian*	*octombrie 2015 – octombrie 2016*
7.	*Activitatea 3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti*	*Cucu Horia; Buzo Andi; Burileanu Corneliu; Banica Alina; Minca Florentina*	*iunie 2016 – august 2016*
8.	*Activitatea 3.8. Dezvoltarea unor modele acustice multilingve si pentru limba engleza (partea I)*	*Cucu Horia; Buzo Andi; Burileanu Corneliu; Banica Alina; Minca Florentina*	*iulie 2016 – octombrie 2016*

3 Rezumat activități

3.1 A3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)

Această activitate a constat in segmentarea si anotarea corpusului de vorbire inregistrat in etapa anterioara (2/2015). Aceste operatii sunt absolut necesare pentru obtinerea unei resurse lingvistice utile in sinteza de vorbire pornind de la text.

3.2 A3.2 Validarea manuală a corpusului de vorbire

In aceasta activitate RACAI a efectuat verificarea corectitudinii alinierilor si a anotarilor facute automat. Astfel, au fost verificate manual 64 fisiere selectate in mod aleatoriu, 32 din corpusul masculin, 32 din cel feminin, dintr-un total de 1000 pentru fiecare vorbitor.

3.3 A3.3 Implementarea unui controller de voce (partea a II-a)

Controllerul de voce este responsabil pentru intreaga logica (software) a aplicatiei. In scenariul standard, controllerul de voce (CV) primeste o comanda vocala de la utilizator, o trimite mai departe catre sistemul de

recunoastere de voce (ASR) pentru a obtine transcrierea vocii. Textul obtinut este procesat astfel incat sa fie inteles de sistem, pentru a decide ce comanda trebuie trimisa mai departe echipamentelor inteligente, folosind standardul KNX. In final, CV decide ce raspuns sa sintetizeze ca feedback pentru utilizator in urma comenzii si a primirii unui status de la echipamentul catre care s-a trimis comanda (daca nu este necesara asteptarea unui feedback de confirmare a executarii actiunii respective, atunci se poate sintetiza un raspuns folosind sistemul TTS pe loc). Este de asemenea posibil sa programam sistemul CV sa porneasca automat o interactiune cu utilizatorul. De exemplu, utilizatorul se intoarce acasa si este intrebat de sistem daca sa seteze temperatura la o anumita valoare. Indiferent de cine initiaza procesul de interactiune, comunicarea va fi incheiata printr-un mesaj sintetizat.

In cadrul acestei activitati a fost implementat controlerul de voce pe un dispozitiv mobil, integrand serviciul RACAI de sinteza de voce (TTS), serviciul UPB de recunoastere de voce (ASR) intr-un controller care proceseaza comenzile primite si le transmite dispozitivelor inteligente atasate prin standardul industrial KNX.

3.4 A3.4. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea a II-a)

În cadrul activității 3.4 (Implementarea parametrilor PNCC şi a tehnicilor de reducere a zgomotului (partea a II-a)), este continuată direcția de cercetare începută în partea I, asupra modulului extragere a parametrilor vocali de tip PNCC (Power Normalized Cepstral Coefficients). Aceşti parametrii, robuşti la zgomot,au fost implementați în CMU Sphinx (utilitar open-source de recunoaştere a vorbirii) şi reprezintă nucleul sistemului SpeeD de recunoaştere automată a vorbirii (LiveTranscriber Server).

În prima parte a cercetării am evidențiat faptul că aceşti tip de parametrii vocali oferă rezultate mai bune, în condiții de laborator, pentru vorbirea ce conține zgomot, în cadrul evaluărilor în care s-au realizat transcrierile de test. În continuare, în partea a doua a implementării, au fostantrenate o serie de modele acustice cu parametrii de tip PNCC, ce au fost ulterior testate pe bazele de date de vorbire achiziționate în condiții reale (homeAuto2 / homeAuto3), în cadrul casei inteligente DOMUS, deținută de Laboratorul de Informatică din Grenoble.

3.5 A3.5. Compararea si evaluarea performantelor tehnicilor implementate

În cadrul activității 3.5 (Compararea şi evaluarea performanțelor tehnicilor implementate) sunt realizate numeroase teste şi evaluări asupra bazelor de date audio achiziționate în condiții reale (homeAuto2 / homeAuto3), în cadrul casei inteligente Domus, deținută de Laboratorul de Informatică din Grenoble. Scopul acestei activități a fost de a determina validitatea folosirii unor modele acustice de tip PNCC, în defavoarea celor ce folosesc parametrii clasici de tip MFCC, precum şi îmbunătățirile aduse prin ajustarea gramaticii cu stări finite (FSG). Pentru a ilustra robustețea sistemului RAV cu bazele de date achiziționate, acest subcapitol ilustrează câştigurile relative ale modelelor PNNC, antrenate cu aceleaşi date ca şi cele de tip MFCC, precum şi configurația ce oferă cele mai bune metrici în situații cât mai apropiate de condițiile reale, în cadrul unei case inteligente.

3.6 A3.6. Diseminarea rezultatelor proiectului (partea I)

O parte din rezultatele proiectului au fost diseminate în cinci articole ştiințifice, dintre care: un articol într-o revistă indexată Thomson Reuters (ISI), două articole în conferințe indexate Thomson Reuters (CPCI, ex ISI), un articol într-o revistă indexată în baze de date internaționale (Google Scholar) şi o conferință indexată in baze de date internaționale (ACM Digital Library).

De asemenea, proiectul ANVSIB a fost prezentat de către Alina Bănică în Laboratorul de cercetare LIMSI, Paris, Franța, în cadrul stagiului de cercetare pe care l-a efectuat în perioada octombrie – noiembrie 2015.

În vederea promovării şi diseminării rezultatelor proiectului ANVSIB, IWAVE a participat în perioada 20-22 octombrie 2016 la targul „ROMANIAN SECURITY FAIR 2016”[1].

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.

[1] ROMANIAN SECURITY FAIR 2016: http://www.romaniansecurityfair.eu/ro

3.7 A3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti

În cadrul activității 3.7 (Studiul algoritmilor state-of-the-art în detecția multilingva a termenilor vorbiți), a fost realizat un raport al stării artei în domeniul detecției şi recunoaşterii multilingve a termenilor vorbiți. Recunoaşterea automată a vorbirii este un domeniu foarte vast, ce absoarbe multe direcții de cercetare din domeniul prelucrării digitale a semnalului vocal. Disponibilitatea online a unor cantități mari de date (corpusuri), a permis utilizarea unor modele statistice la orice nivel al unei arhitecturi RAV, de la partea acustică la modelarealingvistică. Disponibilitatea acestor modele, a tehnicilor moderne de „învățare" a maşinilor (Machine Learning), precum şi succesul comercial al dispozitivelor ce folosesc vocea ca interfață cu utilizatorul, au acceleratdezvoltarea metodelor şi algoritmilor acestor sisteme de recunoaştere, şi au dus la apariția unor noi nişe în cadrul recunoaşterii automate a vorbirii: recunoaşterea multilingvă cu detecția automată a limbii vorbite, recunoaşterea audio multilingvă a termenilor vorbiți, extragerea unor termeni audio din baze de dateaudio multilingve, etc.

Astfel, doar o simplă recunoaştere a vorbirii, într-o limbă cunoscută, nu mai este de ajuns pentru a cataloga un dispozitiv sau o aplicațieca fiind „inteligentă". Utilizatorii aşteaptă mai multă interacțiune de laacestea, cum ar fi Google Now, Apple Siri, Microsoft Cortana, ce pot oferi în timp real răspunsul la întrebări relativ complexe în limbi multiple, accesând în spate conținut text sau audio în diverse limbi (baze de date multilingvistice). Multilingvismul este o caracteristică a acestor aplicații ce implică utilizarea, în mod nativ, a mai mult de o limbă în procesul de recunoaştere. În ziua de azi, aceste caracteristici sunt din ce în ce mai necesare oricărui asistent digital ce se bazează pe ASR în procesul de decodare şi înțelegere a mesajului audio, rostit de către utilizator, în multiple limbi.

Însă înainte ca un sistem ASR să înceapă procesarea şi decodarea documentelor audio în formă brută, limba vorbită trebuie mai întâi identificată, aici fiind vorba de sistemele de „Identificare a limbii vorbite", un subdomeniu al ASR încă nerezolvat complet, cel puțin nu pentru limbile unde resursele sunt limitate (aşa zisele limbi slab dotate, unde nu există destule date pentru a construi modele acustice sau de limbă robuste). În aceste cazuri, comunitatea ştiințifică a încercat o abordare de tip zero-resurse, prin căutarea directă în spațiul audio, în mod nesupervizat, fără sau cu informații limitate de la modelele statistice existente. Aceste cazuri reprezintă o nişă a ASR, un sub-domeniu denumit detecția multilingvă a termenilor vorbiți, fiind detaliat, pe larg, în cadrul studiului.

3.8 A3.8. Dezvoltarea unor modele acustice multilingve si pentru limba engleza (partea I)

În cadrul activității 3.8 au fost realizate experimente ce au avut ca scop crearea unor modele acustice atât pentru limba engleză, cât şi pentru limba română. Corpusul audio a fost împărțit în seturi de antrenare şi testare. Au fost utilizate modele de limbă probabilistice de tip 3-gram, cel în limba engleză fiind disponibil online, iar cel în limba română a fost creat anterior în cadrul grupului de cercetare SpeeD. Pentru antrenarea modelelor acustice s-a folosit în special utilitarul Kaldi, ce oferă unele facilități suplimentare față de CMU Sphinx. Pentru fiecare dintre cele două limbi au fost create şi evaluate câte 5 tipuri de modele acustice, ce diferă între ele prin algoritmii pe care se bazează (platforma HMM-GMM, respectiv rețele neuronale) şi prin diverse adaptări specifice sistemelor de recunoaştere automată a vorbirii, ce se gasesc implementate în utilitarul Kaldi. Evaluarea modelelor acustice obținute a oferit informații cu privire la rata de eroare la nivel de cuvânt. De fiecare dată, s-a observat că modelele antrenate cu rețele neuronale sunt mult mai performante în comparație cu celelalte.

În încheierea secțiunii sunt prezentate două abordări posibile pentru dezvoltarea în viitor a unor modele acustice multilingve.

4 Descrierea ştiințifica şi tehnică, cu punerea în evidență a rezultatelor activităților si gradul de realizare a obiectivelor

4.1 A3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)

Această activitate a constat in anotarea corpusului de vorbire inregistrat in etapa anterioara (2/2015). 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 au fost prezentate in RST-ul precedent.

In aceasta etapa am finalizat corpusul obtinand atat corpusul cu inregistrari audio avand aliniera si anotarea fonetica efectuata, cat si modelele de sinteza de voce antrenate pe corpus. De mentionat ca avem doi vorbitori, o voce feminina si una masculina, rezultand doua corpusuri anotate si doua modele de voce distincte.

4.2 A3.2 Validarea manuală a corpusului de vorbire

În cadrul acestei activități, RACAI a efectuat verificarea corectitudinii alinierilor si a anotarilor facute automat. Astfel, au fost verificate manual 64 fisiere selectate in mod aleatoriu, 32 din corpusul masculin, 32 din cel feminin.

Verificarea primara a fost facuta in vederea stabilirii corectitudinii alocarii fonemelor (deoarece fisierele au trecut initial prin procedeul de transcriere fonetica). Au fost detectate erori minore, precum a fost prezentat in RST-ul precedent. Aceste erori tin de transcrierea automata care uneori aloca alte foneme in loc de cele corecte. Totusi, erorile sunt neglijabile (ex: marcare ‚i' lung vs ‚i' scurt).

In continuare au fost verificate corectitudinea alinierii in timp a fonemelor. Acest lucru s-a efectuat folosind programul Audacity (free, open-source), incarcand un fisier wav si trecand, rand pe rand peste fiecare cuvant (start-stop), apoi, prin sampling, s-a incercat ascultarea unei foneme (acolo unde a fost posibil, deoarece durata unei foneme este de ordinul milisecundelor – am selectat cuvinte simple, cu foneme clar identificabile). Am observat ca alinierea a fost efectuata corect, fara decalaje. Bineinteles, exista o marja de eroare, deoarece in anumite cazuri nu este clar unde ar trebui sa porneasca sau sa se incheie un cuvant. De exemplu, chiar daca inregistrarile s-au efectuat in mediu controlat, exista cazuri in care vorbitorul „uneste" cuvinte diferite prin „mmm" sau „aaa/ăăă", aici fiind usor neclar in ce milisecunda incepe sau se termina un cuvant. Cu exceptia acestor cazuri, verificarea nu a scos la iveala erori in alinierea automata.

In final, am urmarit procesul de anotare automata. In cazul unor feature-uri, este simplu de verificat corectitutdinea anotarii pentru, de exemplu, pozitia in propozitie sau accentul pe silaba; pe de alta parte, verificarea partii de vorbire implica sa avem propozitia in fata, anotata la nivel de cuvant cu parte de vorbire, si sa analizam acolo unde au fost efectuate marcaje gresite. Aceasta verificare a fost facuta doar partial doarece performantele etichetatorului morfo-sintactic automat au fost analizate pe larg in publicatiile anterioare [Boros 2013].

In concluzie, aceasta activitate a determinat acuratetea etichetarii morfosintactice de 97%, a alinierii in timp la nivel de fonem de 98% (exceptand cazurile tip „mmm", tinand cont totusi ca aceasta estimare este subiectiva, neexistand un standard). Restul de anotari fiind efectuate in procend de 100% corect. Corpusul respecta standardele necesare folosirii in antrenarea unui model de sinteza performant.

4.3 A3.3 Implementarea unui controller de voce (partea a II-a)

Controllerul de voce este responsabil pentru intreaga logica (software) a aplicatiei. In scenariul standard, controllerul de voce (CV) primeste o comanda vocala de la utilizator, o trimite mai departe catre sistemul de recunoastere de voce (ASR) pentru a obtine transcrierea vocii. Textul obtinut este procesat astfel incat sa fie inteles de sistem, pentru a decide ce comanda trebuie trimisa mai departe echipamentelor inteligente, folosind standardul KNX. In final, CV decide ce raspuns sa sintetizeze ca feedback pentru utilizator in urma comenzii si a primirii unui status de la echipamentul catre care s-a trimis comanda (daca nu este necesara asteptarea unui feedback de confirmare a executarii actiunii respective, atunci se poate sintetiza un raspuns folosind sistemul TTS pe loc). Este deasemenea posibil sa programam sistemul CV sa porneasca automat o interactiune cu utilizatorul. De exemplu, utilizatorul se intoarce acasa si este intrebat de sistem daca sa seteze temperatura la o anumita valoare. Indiferent de cine initiaza procesul de interactiune, comunicarea va fi incheiata printr-un mesaj sintetizat.

Arhitectura sistemului permite integrarea unui numar arbitrar de sisteme TTS, ASR sau CV. Sistemul CV a fost implementat ca o aplicatie de sine-statatoare, ruland pe sistemul Android. Acest lucru permite o flexibilitate sporita si scade impactul asupra serverelor, in sensul ca procesarea textului (NLP) este

distribuita instantelor individuale de CV, care, in acelasi timp, ofera si interfata spre utilizator. In sarcina serverelor ramane in continuare recunoasterea si sinteza de voce. Aplicatia va „locui” in sisteme special constuite, ruland Android, fiind incastrate in peretii casei (sau alte puncte dorite de utilizator). Deasemenea, aplicatia poate fi instalata si pe telefonul utilizatorului, oferindu-i in acest caz si o interfata grafica, alaturi de posibilitatea de a controla casa de la distanta prin intermediul conectivitatii dispozitivului sau la internet.

Algoritmul de procesare de text al CV este bazat pe o serie de reguli predefinite la instalarea sistemului in casa inteligenta. Setul de reguli este adaptat utilizatorului si a dispozitivelor instalate in casa. Aceste reguli poti fi editate oricand de catre utilizator (de exemplu la instalarea unui nou dispozitiv ce extinde setul de actiuni posibile in casa). Acest lucru se efectueaza modificand un fisier de configurare. In prezent, automatizarea se bazeaza pe reguli de gramatica, care pot fi modificate intr-un mod rapid si simplu. In viitor dorim extinderea suportului pentru alte limbi si integrarea cu alte sisteme anexe, precum organizatoare, calendare, sisteme e-mail, vreme sau roboti conversationali (sisteme intrebare-raspuns).

In continuare prezentam detaliile sistemului CV, apelat prin numele „Casandra”.

4.3.1 Livrabil D3.12 - Model functional pentru controlerul de voce

Interfața WEB pentru configurare (Cassandra web interface)

Aceasta componentă a sistemului este destinată configurării şi descrierii modului in care utilizatorul va interacționa cu sistemul inteligent de control al casei. O „interacțiune”, în contextul sistemului prezentat, defineşte un tip de sarcină pe care sistemul de automatizare al casei trebuie sa o indeplinească în urma unei comenzi vocale primită de la utilizator. Prin intermediul acestei interfețe web, oricărui utilizator inregistrat ii este permis să definească interacțiuni noi, să şteargă şi să modifice interacțiuni deja existente. Pentru fiecare dintre interacțiuni, este necesar să fie introduse exemple de propoziții pe care sistemul de recunoaştere automată a vorbirii să le interpretezedrept comenzi pentru respectiva interacțiune, dar sipropoziții ce vor fi sintetizate ca răspuns al sistemului după procesarea comenzii. De asemenea, o comandă KNX trebuie configurată pentru orice interacțiune. Ea trebuie construită după modelul din imaginea de mai jos: „objectName” primeşte ca valoare un nume sugestiv pentru acțiune, „valueName” primeşte „value”, iar „value” primeşte „true” / „false” dacă interacțiunea estetip „activează”/ „dezactivează”, „deschide”/ „ închide” sau „$1” dacă interacțiunea este de tip „setează”.

```
{
"objectName":"temp",
"valueName":"value",
"value":$1
}
```

Figura 1 Exemplu de comandă KNX

De exemplu, pentru interacțiunea „Temperatura” putem folosi comenzi cum ar fi: „Setează temperatura pe $1.” sau „Vreau să faci $1 grade în sufragerie” iar ca raspuns putem seta „Temperatura este ${”objectName”:”temp”,”valueName”:”value”}”. Se va folosi „$1”ca parametru pentru toate exemplele de comenzi care implică transmiterea de valori către KNX.

Modul de funcționare

Toate datele introduse în interfața web pentru configurare sunt folosite pentru a crea modelul de functionare al sistemului de automatizare a casei.

Pe baza exemplelor de propoziții introduse la configurare se construiesc două seturi de antrenare ambele folosite pentru a antrena doi clasificatori ID3 (metoda prezentata in sectiunea urmatoare). Modelul generat de primul clasificator este folosit pentru a prezice interacțiunea determinată de o comandă, iar celălalt model pentru a determina parametrii dintr-o comandă.

În corpusul folosit pentru determinarea interacțiunilor s-au folosit ca atribute toate cuvintele dintr-o propoziție şi precedența lor codificată astfel:pentru fiecare cuvant „cuvant<cuvant_urmator”, pentru toate cuvintele ce urmează după cel curent. Un exemplu din acest corpus arata astfel: *„temperatura _<seteaza seteaza seteaza<temperatura seteaza<pe temperatura temperatura<pe pe”* .Eticheta se afla pe prima poziție şi este interacțiunea pentru care a fost definită propoziția la configurare.În figura urmatoare este reprezentata o parte din arborele de decizie construit de algoritmul ID3 pe baza acestui corpus. Se observă că dacă un exemplu contine „aerul” şi „casandra porneste”, acesta sigur se referă la interacțiunea„ac_on”.

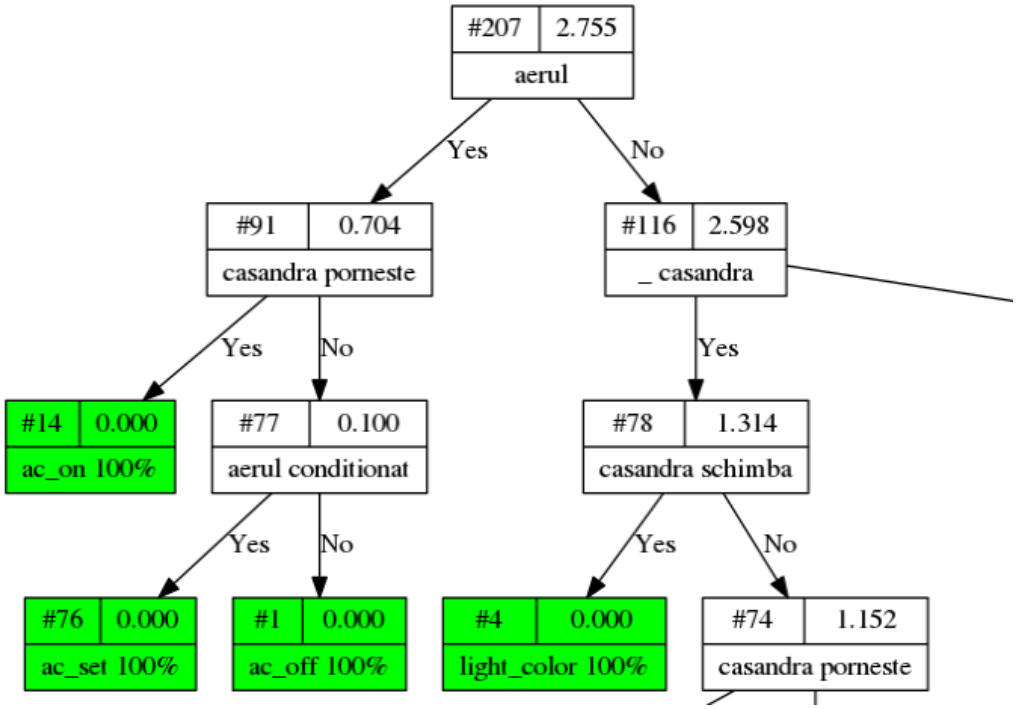

Figura 2 Arbore parțial de decizie ce determină interacțiunea, construit cu algoritmul ID3

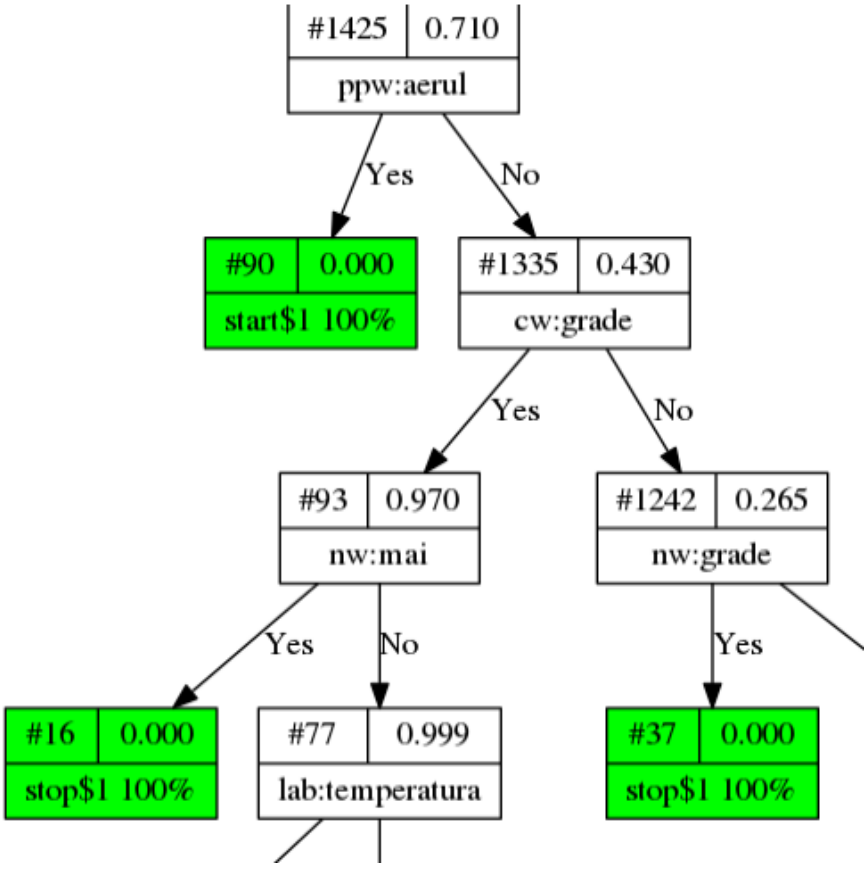

Figura 3 Arbore parțial de decizie ce determină parametrii dintr-o comandă construit cu algoritmul ID3

Pentru cel de-al doilea corpus, folosit pentru a extrage parametrii dintr-o comandă, a fost construit un exemplu de antrenare pentru fiecare cuvânt care nu este parametru. Intrările au una dintre următoarele forme: „_ lab:temperatura ppw:_ pw:_ cw:seteaza” , „stop$1 lab:temperatura nw:in nnw:sufragerie cw:grade” sau„start$1 lab:temperatura ppw:vreau pw:sa cw:faci”. Pe prima poziție se află etichetele codificate astfel : „start$1” dacă după cuvântul curent urmează un parametru, „stop$1” în situația în care cuvântul precedent celui curent este parametru sau „_” când nu ne aflăm în niciuna dintre situațiile de mai sus. Atributele folosite sunt următoarele: „ppw:” – previus previous word, „pw:” – previous word , „cw:” – current word,

„nx:” – next word, „nnx:” – next next word şi „lab:” – interacţiunea din care face parte comanda (etichetă din primul corpus). Pentru cuvintele ce se afla înainte de un parametru s-au folosit doar „ppw:” si „pw:” iar pentru cele care se afla după un parametrus-aufolositdoar „nw:” şi „nnw:”. Mai jos, este prezentat şi pentru acest set de date, un arbore de decizie parțial construit cu algoritmul ID3. Din acest arbore reiese că dacă unul dintre exemple conține atributul „ppw:aerul” atunci sigur cuvântul următor celui curent este un parametru, iar dacă nu conține acest atribut dar conține „cw:grade” si „nw:mai” atunci cu siguranță cuvântul precedent celui curent este un parametru.

Modele obținute după antrenarea celor doi clasificatori pe baza seturilor de date create sunt folosite pentru a interpreta comenzile primite de la utilizator.

Algoritmul ID3 a fost implementat in Java si este impachetat ca o librarie de sine statatoare, anexata proiectului ANVSIB. Am antrenat algoritmul ID3 generand intre 15-100 propozitii parametrizate. Acuratetea obtinuta pe setul de train este de 94.2-99.8%.

4.4 A3.4. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea a II-a)

În ultimul timp, o nouă tehnică de extragere a parametrilor caracteristici ai semnalului vocal, denumită coeficienţi PNCC (Power Normalized Cepstral Coefficients), a apărut. Aceasta implică folosirea unor bancuri de filtre de tip gammatone, în locul celor triunghiulare. În teorie, ar trebui să ofere o precizie de recunoaştere superioară în condiţii de zgomot şi reverberaţie, folosind aproximativ aceleaşi resurse de computaţie, comparabile cu cele ale MFCC. Următoareasecţiune va face o comparaţie a performanţelor acestor parametri vocali, în sarcina de recunoaştere la distanţă, în cadrul unei case inteligente.

Coeficienţii PNCC pot fi consideraţi ca o variantă a parametrilor MFCC, dar cu diferite etape ale algoritmului convenţional înlocuite de procese inspirate de sistemul auditiv uman. Se aseamănă cu alţi parametrii clasici (PLP - Coeficienţii cepstrali de predicţie liniară), înlocuind răspunsul neliniar al funcţiei logaritmice cu o funcţie neliniară proporţională cu puterea semnalului analizat, integrând în acelaşi timp o serie de filtre gammatone. Un atribut cheie al acestor parametri este că ei folosesc a) analiza de putere pe termen mediu, în care parametrii de mediu sunt estimaţi pe o durată de timp mai mare decât cea utilizată în mod frecvent pentru procesarea vorbirii, b) frecvenţa de uniformizare şi c) un algoritm de reducere a zgomotului bazat pe o filtrare asimetrică ce elimină zgomotul de fond. Diagrama bloc a extragerii parametrilor PNCC este expusă în Figura 4.

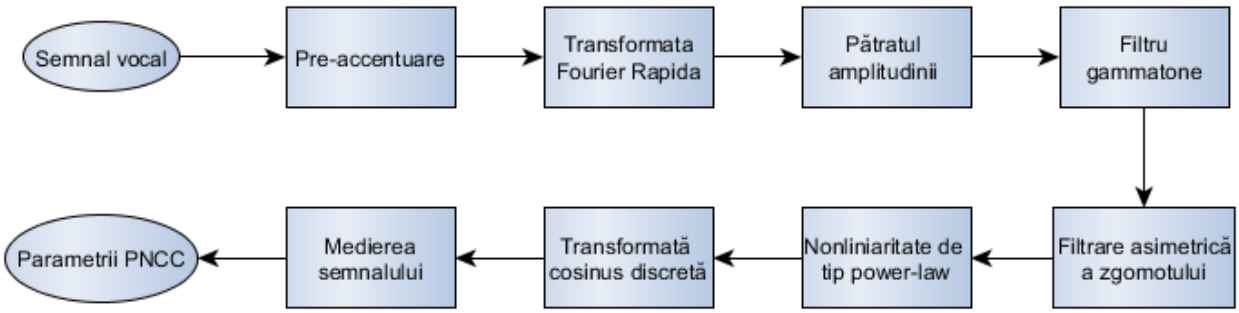

Figura 4 Diagrama bloc a extragerii parametrilor de tip PNCC

Având în vedere cele de mai sus, aceşti parametrii au fost implementaţi în cadrul părţii I al acestei activităţi şi evaluaţi folosind baze de date achiziţionate de grupul de cercetare SpeeD, înainte de începerea acestui proiect, după cum urmează:

- RSC-eval (read-speech corpus), un corpus de vorbire continuă, citită, în limba română, cuprinzând înregistrări curate (fără zgomot);
- SSC-eval (spontaneous-speech corpus), un corpus ce conţine vorbire spontană înregistrată şi în condiţii de linişte şi în condiţii de zgomot (zgomot sau muzică de fundal, vorbire la telefon, etc).

Evaluarea a fost realizată cu ajutorul unui model acustic antrenat cu parametrii vocali PNCC şi un model acustic antrenat cu parametrii vocali MFCC, zgomotul fiind adăugat digital la diferite rapoarte semnal-zgomot.Aceste rezultate preliminare au relevat faptul că, indiferent de tipul de zgomot şi SNR, modelul acustic PNCC este mai bun ca cel de tip MFCC, reducerea relativă WER fiind, în medie, între 10% şi 19%, în funcţie de SNR.

Având în vedere cele de mai sus, în partea a doua a implementării, au fost antrenate o serie de modele acustice cu parametrii de tip PNCC, ce au fost ulterior testate pe bazele de date de vorbire achiziționate în condiții reale,în Laboratorul de Informatică din Grenoble, de către membrii echipei SpeeD.Bazele de date au fost achiziționate în două etape, având următoarele denumiri: homeAuto2 şi homeAuto3. Bazele de date conțin 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 de comenzi în aşa fel încât să acopere diferite moduri de a exprima aceeaşi comandă,iar fiecare comandă a fost rostită de un anumit număr de ori, în funcție de frecvența 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 distanțe față de microfoane.

Tabelul 1 prezintă modelele acustice de tip MFCC folosite în cadrul comparației şi evaluării realizate în această parte a doua a activității, precum şi modelele noi antrenate de tip PNCC, folosind modulul de extragere a parametrilor PNCC implementat în cadrul acestui proiect, în nucleul sistemului ASR SpeeD (LiveTranscriber server).

Tabel 1 Modelele acustice folosite în cadrul evaluării tehnicilor implementate de reducere a zgomotului

Nume model acustic	**branch1**	**mixModels**	**spk01-20**	**AM002**	**AM001**	**AM010**	**AM053**
Durată (ore)	87	73	40	127	127	100	227
# vorbitori	200+	70	17	200+	200+	200+	200+
Tip vorbire	dictare + vorbire spontană	dictare	dictare	dictare + talkshows	talkshows + vorbire spontană	dictare + vorbire spontană	talkshows + dictare + vorbire spontană
# senone	2000	4000	4000	4000	4000	4000	4000
# densit Gauss	16	64	32	128	128	64	128
# stări / HMM	3	3	3	3	3	3	3
Tip parametrii	MFCC	MFCC	MFCC	MFCC	PNCC	PNCC	PNCC

Pentru a obține cel mai bun WER pentru modelele de tip PNCC, a fost realizată şi o primă etapă de calibrare (tuning) a următorilor parametrii de decodare: Word Insertion Probability (WIP), Out-of-Grammar Probability (OOGG), respectiv Relative Beam Width (RBW). Valorile obținute au fost sumarizate în tabelele 2, 3 şi 4, valorile optime alese fiind cele ce oferă cel mai mic SER (Sentance Error Rate - Rata de eroare la nivel de propoziție).

Tabel 2 Calibrarea parametrilor RBW şi OOGP pentru AM001

					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
RBW Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**
	85	10	2	30	11.3	25.1	233.0	100.0
	85	10	2	40	8.8	21.2	252.9	100.0
	85	10	2	50	8.2	20.5	200.0	75.7
	85	10	2	60	8.0	19.9	102.4	34.8
	85	10	2	70	7.1	19.1	24.8	7.6
	85	10	2	80	6.8	18.8	24.7	7.5
	85	10	2	90	6.9	18.9	17.6	4.8
	85	10	2	100	6.8	18.9	17.6	4.8
					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
OOGP Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**

	2	10	2	70	10.0	20.7	4.8	1.4
	5	10	2	70	9.8	20.6	4.8	1.4
	10	10	2	70	9.4	20.3	6.7	1.9
	20	10	2	70	8.2	19.8	11.4	2.9
	30	10	2	70	7.9	19.6	12.4	3.3
	40	10	2	70	7.5	19.3	24.8	7.6
	50	10	2	70	7.3	19.1	107.9	35.2
	60	10	2	70	7.1	19.1	228.8	77.1
	70	10	2	70	7.1	19.1	294.8	100.0
	80	10	2	70	7.1	19.1	294.8	100.0

Tabel 3 Calibrarea parametrilor RBW și OOGP pentru AM010

					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
RBW Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**
	85	10	2	30	14.4	28.9	221.4	100.0
	85	10	2	40	9.7	23.0	237.6	100.0
	85	10	2	50	8.4	21.5	194.3	75.7
	85	10	2	60	8.1	20.9	110.0	37.1
	85	10	2	70	7.6	20.7	12.9	4.3
	85	10	2	80	7.7	20.6	7.1	1.9
	85	10	2	90	7.6	20.8	7.1	1.9
	85	10	2	100	7.8	20.8	8.1	1.9
					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
OOGP Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**
	2	10	2	70	20.1	29.5	0.0	0.0
	5	10	2	70	19.3	29.3	4.8	0.5
	10	10	2	70	18.3	28.5	1.8	0.5
	20	10	2	70	15.3	26.4	1.9	0.5
	30	10	2	70	11.7	23.2	3.8	1.0
	40	10	2	70	10.7	22.7	12.6	4.3
	50	10	2	70	9.3	21.7	114.8	37.1
	60	10	2	70	7.8	20.8	211.4	75.2
	70	10	2	70	7.6	20.7	275.7	100.0
	80	10	2	70	7.6	20.7	269.3	100.0

Tabel 4 Calibrarea parametrilor RBW și OOGP pentru AM053

					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
RBW Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**
	85	10	2	30	10.0	23.0	202.9	100.0
	85	10	2	40	8.3	20.5	223.8	100.0
	85	10	2	50	7.4	19.7	201.4	84.3

	85	10	2	60	19.2	19.2	110.5	39.0
	85	10	2	70	18.8	18.8	22.4	7.1
	85	10	2	80	18.9	18.9	22.4	7.1
	85	10	2	90	19.0	19.0	24.3	7.1
	85	10	2	100	19.1	19.1	27.1	8.1
					WER [%] on in-grammar utterances	**SER [%] on in-grammar utterances**	**WER [%] on out-of-grammar utterances**	**SER [%] on out-of-grammar utterances**
OOGP Tunning	**OOGP (1E-)**	**WIP (1E-)**	**LW**	**RBW (1E-)**	**homeAuto3 cu comenzi**	**homeAuto3 cu comenzi**	**homeAuto3 fara comenzi**	**homeAuto3 fara comenzi**
	2	10	2	70	7.7	19.3	5.7	1.9
	5	10	2	70	7.5	19.1	5.6	1.9
	10	10	2	70	7.5	19.1	8.6	2.9
	20	10	2	70	6.9	18.8	11.9	3.8
	30	10	2	70	6.7	18.8	17.6	5.7
	40	10	2	70	6.6	18.8	22.4	7.1
	50	10	2	70	6.5	18.8	118.1	40.5
	60	10	2	70	6.5	18.8	223.6	85.7
	70	10	2	70	6.5	18.8	267.1	100.0
	80	10	2	70	6.5	18.8	267.1	100.0

În timpul etapei de calibrare, au fost aduse numeroase îmbunătăţiri şi gramaticii cu stări finite (FSG) folosite de modulul ASR al laboratorului SpeeD (Anexa 1). Îmbunătăţirile la modelul FSG au adus un câştig relativ de aproximativ 5-7% la metrica SER, în special asupra bazei de date homeAuto3, ce cuprinde mai mulţi vorbitori şi mai multe comenzi.

4.5 A3.5. Compararea si evaluarea performantelor tehnicilor implementate

Pentru evaluarea performanţelor sistemului, precum şi a parametrilor de tip PNCC implementaţi, au fost realizate o serie de experimente pe bazele de date achiziţionate direct în cadrul casei inteligente DOMUS (homeAuto2 şi homeAuto3), variind modelele acustice, după etapa de calibrare, între cele clasice, de tip MFCC şi cele de tip PNCC. De asemenea, au fost testate şi modele de limbă statistice, pentru a observa dacă constrângerile impuse de un model de limbă cu gramatică cu stări finite reprezintă cea mai bună opţiune, în cazul rostirii comenzilor în cadrul unei case inteligente.

Tabelul 5 prezintă bazele de date testate şi informaţii legate despre denumirea şi conţinutul lor, ele fiind folosite în cadrul tabelului ce conţine rezultatele experimentale.

Tabel 5 Bazele de date folosite în cadrul evaluării performanţelor tehnicilor implementate

Bază de date	**Nr. vorbitori**	**Descriere**
homeAuto2	5	Baza de date ce conţine doar un număr restrâns de comenzi
homeAuto2-no-139	4	Baza de date homeAuto2 ce nu conţine unul dintre vorbitori, cel mai zgomotos
homeAuto3-allch-allwav	11	Baza de date homeAuto3 ce conţine toate fişierele audio, precum şi toate canalele audio
homeAuto3-MBEST-all	11	Baza de date homeAuto3, conţinând toate fişierele audio, cel mai bun canal
homeAuto3-MBEST-comenzi	11	Baza de date homeAuto3, cel mai bun canal, doar fişierele audio ce conţin comenzi
homeAuto3-MBEST-fara-comenzi	11	Baza de date homeAuto3, cel mai bun canal, doar fişierele fără comenzi

Ca şi metrică principală de evaluare, am decis să folosim DER (Domotic Error Rate), o metrică recent introdusă, special folosită pentru evaluarea erorii în cadrul sistemelor de automatizare a clădirilor inteligente, precum şi a ratei de eroare la nivel de propoziţie, SER (Sentence Error Rate).

Domotic Error Rate (DER) este definită după cum urmează [1]:

$$DER = \frac{Missed + False\ Alarms}{Voice\ Commands_{syntactically\ correct}} \quad (1)$$

Pentru DER, etalonul (ground truth) reprezintă numărul de comenzi rostite ce respectă gramatica. Acele comenzi ce nu respectă gramatică (intenţia vorbitorului a fost de a rosti o comandă, dar a rostit-o greşit), nu sunt luate în considerare. Comenzile ratate (missed) corespund celor ce nu au fost corect recunoscute, iar alarmele false sunt cele clasificate incorect ca şi comenzi vocale.

Metricile componente DER sunt calculate folosind următoarele criterii:

$$\%FA\ (False\ Alarms) = \frac{Nr.\ propoziţii\ transcrise\ greşit - Nr.propoziţii\ cu\ transcrieri <unk> (unknown)}{Total\ propoziţii}, \quad (2)$$

pentru propoziţii ce conţin comenzi, iar pentru cele ce nu conţin comenzi, formula de calcul este:

$$\%FA\ (False\ Alarms) = \frac{Transcrieri\ diferite\ de <unk> (unknown)}{Total\ propoziţii} \quad (3)$$

$$\%MISS = \frac{Nr.propoziţii\ ce\ conţin <unk>\ (unknown)}{Total\ propoziţii}, \quad (4)$$

pentru propoziţii ce conţin comenzi, 0 (zero) pentru cele ce nu conţin comenzi.

Tabel 6 Evaluarea performanţelor tehnicilor implementate

Nr.	Baza de date	Model acustic	FSG/LM	Miss probability (%)	False alarm probability (%)	Correctly recognized (%)	DER (%)	SER (%)
1	homeAuto2	branch1	FSG	21.6	1.57	76.8	23	23.2
2	homeAuto2	mixModels	FSG	19.2	1.67	79.1	20.8	20.9
3	homeAuto2	spk01-20	FSG	32.9	1.96	65.1	34.9	34.9
4	homeAuto2	branch1	LM	0	28.8	71.2	28.8	28.8
5	homeAuto2	mixModels	LM	0	33.4	66.6	33.4	33.4
6	homeAuto2	spk01-20	LM	0	59.2	40.8	59.2	59.2
7	homeAuto2-no-139	branch1	FSG	5.7	2.3	91.9	8	8.1
8	homeAuto2-no-139	mixModels	FSG	3.4	2.3	94.2	5.7	5.8
9	homeAuto2-no-139	spk01-20	FSG	17.6	2.5	79.8	20.2	20.2
10	homeAuto2-no-139	branch1	LM	0	29	71	29	29
11	homeAuto2-no-139	mixModels	LM	0	32.4	67.5	32.4	32.5
12	homeAuto2-no-139	spk01-20	LM	0	65.5	34.5	65.5	65.6
13	homeAuto2	LiveTranscriber (mixModels)	FSG	18.7	1.8	79.4	20.59	20.6
14	homeAuto2	LiveTranscriber (AM001)	FSG	0	6.9	93	6.9	7
15	homeAuto2	LiveTranscriber (AM002)	FSG	19	1.4	79.5	20.4	20.5
16	homeAuto2	LiveTranscriber (AM010)	FSG	0	6.6	93.3	6.6	6.7
17	homeAuto2	LiveTranscriber (AM053)	FSG	0	6.2	93.7	6.2	6.3
18	homeAuto3-allch-allwav	branch1	FSG	60.7	1.7	37.5	62.5	62.5

19	homeAuto3-allch-allwav	mixModels	FSG	85.6	9.29	23.7	76.3	76.3
20	homeAuto3-allch-allwav	spk01-20	FSG	94	11.4	17.4	82.6	82.6
21	homeAuto3-allch-allwav	branch1	LM	0	98.9	1.1	98.9	98.9
22	homeAuto3-allch-allwav	mixModels	LM	0	99.6	0.4	99.6	99.6
23	homeAuto3-allch-allwav	spk01-20	LM	0	100	0	100	100
24	homeAuto3-MBEST-all	branch1	FSG	13.6	0.9	85.4	14.6	14.6
25	homeAuto3-MBEST-all	mixModels	FSG	35.5	2.1	66.6	33.3	33.4
26	homeAuto3-MBEST-all	spk01-20	FSG	68.1	8	39.9	60	60.1
27	homeAuto3-MBEST-all	branch1	LM	0	87.7	12.2	87.7	87.8
28	homeAuto3-MBEST-all	mixModels	LM	0	96.1	3.9	91.1	96.1
29	homeAuto3-MBEST-all	spk01-20	LM	0	99.8	0.2	99.8	99.9
30	homeAuto3-MBEST-comenzi	branch1	FSG	1.8	13.8	84.3	15.7	15.7
31	homeAuto3-MBEST-comenzi	mixModels	FSG	26.4	11	62.5	37.4	37.5
32	homeAuto3-MBEST-comenzi	spk01-20	FSG	63.7	4	32.2	67.7	67.8
32	homeAuto3-MBEST-fara-comenzi	branch1	FSG	2.38	0	97.5	2.3	2.5
34	homeAuto3-MBEST-fara-comenzi	mixModels	FSG	0	0	0	0	0
35	homeAuto3-MBEST-fara-comenzi	spk01-20	FSG	0	0	0	0	0
36	homeAuto3-MBEST-comenzi	LiveTranscriber (mixModels)	FSG	20.9	13.2	65.8	34.1	34.2
37	homeAuto3-MBEST-fara-comenzi	LiveTranscriber (mixModels)	FSG	0	0	100	0	0
38	homeAuto3-MBEST-comenzi	LiveTranscriber (AM001)	FSG	0.35	18.5	81.1	18.8	18.9
39	homeAuto3-MBEST-fara-comenzi	LiveTranscriber (AM001)	FSG	0	4.7	95.2	4.7	4.8
40	homeAuto3-MBEST-comenzi	LiveTranscriber (AM002)	FSG	2.3	18	79.6	20.3	20.4
41	homeAuto3-MBEST-fara-comenzi	LiveTranscriber (AM002)	FSG	2.3	0	97.5	2.3	2.5
42	homeAuto3-MBEST-comenzi	LiveTranscriber (AM010)	FSG	0	20.6	79.4	20.6	20.6
43	homeAuto3-MBEST-fara-comenzi	LiveTranscriber (AM010)	FSG	0	1.9	98.1	1.9	1.9
44	homeAuto3-MBEST-comenzi	LiveTranscriber (AM053)	FSG	2.1	5.2	92.5	7.4	7.5
45	homeAuto3-MBEST-fara-comenzi	LiveTranscriber (AM053)	FSG	0	2.8	97.1	2.8	2.9

Tabelul 6 sumarizează toate evaluările realizate în cadrul acestei activităţi, în vederea comparării performanţelor sistemului. Acesta indică faptul că modelele acustice de tip PNCC sunt mai bune decât cele clasice, de tip MFCC, în aceleaşi condiţii şi baze de date de test. Pentru baza de date homeAuto2, cel mai bun SER, de 6.4%, este obţinut de modelul acustic de tip PNCC, AM053, cu o gramatică cu stări finite (FSG, linia 17 în Tabelul 6). Ca şi comparaţie, cel mai bun model de tip MFCC pentru homeAuto2 este mixModels (linia 2 în Tabelul 6), cu un SER de 20.9%. În condiţiile eliminării vorbitorilor cu cel mai mult zgomot şi reverberaţie, modelele clasice de tip MFCC pot oferi rezultate similare (linia 8 în Tabelul 6), de 5.8% SER, dar cu un model acustic mixt. Pentru homeAuto3, se păstrează acelaşi trend, modelul acustic de tip PNCC AM053 oferă cele mai bune rezultate, de 7.5% SER pentru o gramatică de tip FSG, faţă de 15.7% SER pentru modelul de tip MFCC branch1.

În concluzie, implementarea modulului de tip PNCC în cadrul aplicației de decodare ASR, a laboratorului SpeeD, a îmbunătățit rezultatele în cazul tuturor bazelor de date achiziționate în cadrul proiectului, oferind o reducere relativă a SER în cazul homeAuto2 de 14.5%, iar pentru homeAuto3 îmbunătățirea fiind de 8.6%, în aceleaşi condiții de test. Aceste rezultate sunt în concordanță cu rezultatele preliminarii, obținute în prima parte a activității, ce au relevat câştigurile obținute prin noul tip de parametru.

4.6 A3.6. Diseminarea rezultatelor proiectului (partea I)

O parte din rezultatele proiectului au fost diseminate în cinci articole ştiințifice, dintre care: un articol într-o revistă indexată Thomson Reuters (ISI), două articole în conferințe indexate Thomson Reuters (CPCI, ex ISI), un articol într-o revistă indexată în baze de date internaționale (Google Scholar) şi o conferință indexată in baze de date internaționale (ACM Digital Library). Referințele complete ale articolelor se găsesc mai jos.

- Mihai Dogariu, Horia Cucu, Andi Buzo, Dragoş Burileanu, Octavian Fratu, "Speech Database Acquisition for Assisted Living Environment Applications," in the Proceedings of the 8th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, 2015, pp. 191-196, ISBN 978-1-4673-7559-7.
- Mihai Dogariu, Horia Cucu, Andi Buzo, Dragoş Burileanu, Octavian Fratu, "Speech Applications in the eWALL Project," in the Proceedings of the 8th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, 2015, pp. 197-204, ISBN 978-1-4673-7559-7.
- Horia Cucu, Andi Buzo, Corneliu Burileanu, "The SpeeD Grammar-based ASR System for the Romanian Language," in Romanian Journal of Information Science and Technology, vol. 18, no. 1, pp. 33-53, 2015, ISSN: 1453-8245.
- Valentin Andrei, Horia Cucu, Lucian Petrică, "Considerations on Developing a Chainsaw Intrusion Detection and Localization System for Preventing Unauthorized Logging," in Journal of Electrical and Electronic Engineering. Vol. 3, No. 6, 2015, pp. 202-207.
- Tiberiu Boroş, Ştefan Daniel Dumitrescu, "Robust deep-learning models for text-to-speech synthesis support on embedded devices," in the Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (MEDES), Caraguatatuba/Sao Paulo, Brazil 2015, pp. 98-102.

De asemenea, proiectul ANVSIB a fost prezentat de către Alina Bănică în Laboratorul de cercetare LIMSI, Paris, Franța, în cadrul stagiului de cercetare pe care l-a efectuat în perioada octombrie – noiembrie 2015. Pe durata acestui stagiu Alina Banica a analizat si modelat secvențele vocalice ale limbilor romanice pentru a înțelege care sunt frecvențele şi proprietățile acestor secvențe care generează erori în sistemele automate de recunoaştere a vorbirii. Analiza s-a realizat pe o bază de date ce conține aproximativ 300.000 de foneme, bază de date obținută din jurnalele de ştiri. Această activitate a fost utilă în crearea unui model fonetic mai robust pentru sistemul de recunoaştere automată a vorbirii.

În vederea promovării şi diseminării rezultatelor proiectului ANVSIB, IWAVE a participat în perioada 20-22 octombrie 2016 la targul „ROMANIAN SECURITY FAIR 2016"[2]. În cadrul targului, IWAVE a participat cu un stand unde a prezentat solutia proof-of-concept Casandra, conectând un dispozitiv mobil la instalatia de iluminare, sistemul de irigatie, sistemul de climatizare si sistemul de alarma. Comenzile au fost date vocal prin intermediul telefonului. Sectiunea Documente in extenso cuprinde poze si filme demonstrative efectuate in cadrul expozitiei.

[2] ROMANIAN SECURITY FAIR 2016: http://www.romaniansecurityfair.eu/ro

În afara celor menționate mai sus, proiectul a fost făcut cunoscut în forumurile ştiinţifice în care laboratorul SpeeD (UPB) are acces. Adresa paginii web a proiectului a rămas neschimbată[3].

4.7 A3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti

Studiul realizat este descris în extenso în documentul ataşat (vezi secţiunea 5 – documente în extenso).

4.8 A3.8. Dezvoltarea unor modele acustice multilingve si pentru limba engleza (partea I)

Unul dintre componentele principale ale unui sistem de recunoaştere automată a vorbirii este modelul acustic. Acesta oferă informaţii despre probabilitatea unui mesaj vorbit, dată fiind secvenţa de cuvinte pronunţată. Construirea unui astfel de model presupune utilizarea unui algoritm de antrenare aplicat pe un corpus de vorbire cât mai mare, pentru care este cunoscută transcrierea text.

Într-o primă fază a experimentelor, pentru crearea modelelor acustice s-a folosit utilitarul CMU Sphinx. Acesta este o platformă de dezvoltare cu sursă deschisă, realizată în cadrul Universităţii Carnegie Mellon, ce cuprinde o serie de algoritmi specializaţi în recunoaşterea automată a vorbirii. Sphinx permite parametrizarea semnalului vocal şi extragerea coeficientilor de tip MFCC, PLP sau LPC. Modelarea acustică presupune crearea unor sisteme statistice de tip HMM-GMM.

Un alt utilitar destul de folosit la crearea sistemelor de recunoaştere a vorbirii este Kaldi. Dezvoltat în cadrul Universităţii Johns Hopkins, Kaldi este o platformă cu sursă deschisă ce oferă unele avantaje faţă de CMU Sphinx. Avantajul major este reprezentat de existenţa algoritmilor de antrenare ce folosesc reţele neuronale adânci (DNN). În urma testelor, s-a observat faptul că aceştia oferă o acurateţe mai bună în comparaţie cu sistemele HMM-GMM. Un alt avantaj este dat de posibilitatea de a aplica o serie de transformate ce conduc la creşterea performanţelor. LDA (Linear Discriminant Analysis) realizează reducerea dimensiunii vectorului de coeficienţi, astfel incât să se pastreze cât mai multă informaţie discriminatorie ce separă clasele fonetice. Coeficienţii sunt apoi decorelaţi prin aplicarea unei transformări liniare simple (MLLT - Maximum Likelihood Linear Transform). SAT (Speaker Adaptive Training) are ca scop creşterea robusteţii sistemului prin adaptarea la vorbitori necunoscuţi sistemului. Un model acustic ce foloseşte SAT are capacitatea de a fi independent de vorbitor. MMI (Maximum Mutual Information) urmareşte să maximizeze probabilitatea aposteriori a secvenţei rostite.

Cu ajutorul utilitarului Kaldi, au fost dezvoltate mai multe modele acustice, atât pentru limba engleză, cât şi pentru limba romană.

Pentru realizarea modelului acustic în limba engleză s-a folosit corpusul de vorbire TED-LIUM, versiunea 1 [Rousseau, 2012], ce conţine înregistrări audio şi transcrierile aferente de la conferinţele TedTalks[4]. Corpusul a fost împărţit în 3 seturi distincte ale căror caracteristici se regăsesc în Tabelul 7.

Tabel 7 Corpusul audio in limba engleză

	Antrenare	Dezvoltare	Testare
Număr de conversaţii	774	19	11
Durata totală • **Vorbitori masculini** • **Vorbitori feminini**	118 h 4 m 48 s 81 h 53 m 7s 36 h 11 m 41 s	4 h 12 m 55 s 3 h 13 m 57 s 58 m 58 s	3 h 4 m 5 s 2 h 21 m 35 s 42 m 29 s
Durata medie	9 m 9 s	13 m 18 s	16 m 43 s
Număr de vorbitori unici	666	19	11

Setul de antrenare a fost folosit la crearea modelului acustic, în timp ce setul de dezvoltare a servit la evaluarea sistemului în vederea ajustării sale. Practic, în funcţie de rezultatele obţinute pe baza setului de dezvoltare, modelul acustic ar putea fi reantrenat cu scopul obţinerii unor rezultate mai bune. Setul de testare a fost folosit la evaluarea finală.

[3] Pagina web a proiectului ANVSIB: http://speed.pub.ro/anvsib
[4] TED: https://www.ted.com

Modelul de limbă utilizat este unul de tip n-gram [Williams, 2015] şi conţine probabilităţile de apariţie pentru: 1-gram: 150000 termeni; 2-gram: 1596991 termeni; 3-gram: 1969287 termeni.

Modelul fonetic conţine transcrierea fonetică a fiecărui cuvânt din vocabular. În acest experiment s-a folosit un număr de 39 foneme verbale şi 7 foneme nonverbale (linişte şi diverse zgomote). Fiecare dintre aceste foneme este tratat diferit în funcţie de poziţia sa în cuvânt (început, sfârşit, interior) şi numărul de apariţii în cuvânt (apariţie unică sau nu).

Parametrii vocali utilizaţi sunt cei de tip MFCC, fiind extraşi vectori cu dimensiunea de 13 coeficienţi, pe baza cărora au fost calculate derivatele de ordin 1 si 2.

Tabel 8 Rezultatele evaluarii pentru limba engleză

Nr. crt	Tipul sistemului acustic	Setul de date	Rezultate WER [%]
1	**TRI1**	**dezvoltare**	31.64
		testare	31.66
2	**TRI2 LDA+MLLT**	**dezvoltare**	27.23
		testare	25.43
3	**TRI3 LDA+MLLT+SAT**	**dezvoltare**	22.56
		testare	20.59
4	**TRI3 MMI**	**dezvoltare**	19.90
		testare	17.94
5	**DNN**	**dezvoltare**	16.60
		testare	15.40

Au fost utilizate mai multe tipuri de modele acustice: TRI1, TRI2 (LDA+MLLT), TRI3 (LDA+MLLT+SAT), TRI3(MMI) si DNN. Primele 4 modele folosesc platforma HMM-GMM, iar fiecare fonem este reprezentat sub forma unui automat finit cu 3 stări, starea iniţială şi finală având rolul de a modela fonemul în raport cu fonemele din imediata sa vecinatate. Cele 4 modele diferă între ele prin îmbunătăţirile ce le-au fost aplicate. Astfel, în cazul setului de testare, utilizarea transformatelor LDA+MLLT a condus la o scadere cu 19.67% a ratei de eroare la nivel de cuvânt (WER). Adaptarea la vorbitor a făcut ca rata de eroare să scadă cu 34.96% faţă de modelul iniţial, iar aplicarea MMI a condus la o scădere cu 43.3%. Ultimul model a fost obţinut prin antrenarea datelor folosind o reţea neuronală. În acest caz a fost înregistrată cea mai mica valoare a erorii, aceasta scăzând faţă de modelul iniţial cu 51.35%.

Pentru relizarea şi evaluarea modelului acustic în limba română s-a folosit un corpus ce conţine vorbire citită în condiţii de linişte, dar şi înregistrări ale unor emisiuni de televiziune şi radio, unele dintre acestea fiind afectate de diverse zgomote de fundal (muzică, râsete, etc.).

Întreg corpusul disponibil a fost împărţit conform Tabelului 9.

Tabel 9 Corpusul audio in limba romana

Set	Subset	Durata		Conţinut
Antrenare	Omega train	94 h 46 m 14 s	225 h 31 m 31 s	Vorbire citita, fara zgomot
	TV train1 (TVDatabase)	27 h 27 m 21 s		Vorbire spontană, uneori cu zgomot
	TV train2 (TVautomaticAquisition)	103 h 17 m 56 s		Vorbire spontană, fără zgomot
Testare	Omega test	5 h 29 m 6 s	8 h 58 m 9 s	Vorbire citită, fără zgomot
	TV test (News all)	3 h 29 m 3 s		Vorbire spontană, uneori cu zgomot

Atât pentru antrenare cât şi pentru testare, au fost folosite subseturi ale corpusului ce diferă prin tipul vorbirii (citită sau spontană) şi prin prezenţa sau absenţa zgomotului de fundal.

Modelul de limbă folosit este unul de tip n-gram creat în cadrul grupului de cercetare SpeeD. Acesta conține probabilitățile de apariție pentru: 1-gram: 64002 termeni; 2-gram: 17939923 termeni; 3-gram: 24912946 termeni.

Tabel 10 Rezultatele evaluarii pentru limba romană

Nr. crt	Tipul sistemului acustic	Rezultate [WER%]	
		Omega test	TV test
1	TRI1	12.69	29.78
2	TRI2 LDA+MLLT	11.73	29.2
3	TRI3 LDA+MLLT+SAT	9.88	28.06
4	TRI3 MMI	9.08	27.03
5	DNN	8.08	24.55

Modelul fonetic conține transcrierea fonetică a fiecărui cuvânt din vocabular. În acest experiment s-a folosit un număr de 36 foneme verbale şi un fonem ce modelează liniştea. Fiecare dintre aceste foneme este tratat diferit în funcție de poziția sa în cuvânt (început, sfârşit, interior) şi numărul de apariții în cuvânt (apariție unică sau nu).

De asemenea, similar cu experimentul pentru crearea sistemului in limba engleză, parametrii vocali sunt de tip MFCC, utilizându-se vectori a câte 13 coeficienți, pe baza cărora au fost calculate derivatele de ordin 1 şi 2.

În urma evaluării sistemului, au fost obținute rezultatele din Tabelul 10.

Primul model este implementat pe baza platformei HMM-GMM, în timp ce următoarele 3 sunt variante îmbunătățite ale acestuia. Rata de eroare la nivel de cuvânt a modelului inițial scade față de următoarele 3 modele cu 7.56%, 22.14% şi 28.44% în cazul evaluării pe setul „Omega test", respectiv cu 1.94%, 22.14% şi 28.44% la evaluarea setului „TV test". Şi de această dată, modelul ce a fost antrenat folosind rețeaua neuronală a obținut cel mai bun rezultat, înregistrându-se o scădere a erorii cu 36,32% față de modelul inițial la evaluarea primului set de test, respectiv 17.56% față de modelul inițial la evaluarea celui de-al doilea set de test.

În viitor, ne propunem crearea unor modele acustice multilingve, ce vor sta la baza unui sistem de recunoaştere automată a vorbirii independent de limba vorbitorului. Pentru realizarea acestei sarcini, sunt posibile doua abordări ale problemei. Prima abordare se referă la crearea unui model acustic unic, la antrenarea căruia se vor folosi corpusuri de vorbire multilingve. Sistemul va fi capabil să realizeze transcrierea vorbirii în oricare dintre limbile cu care a fost antrenat. O provocare în acest caz va fi reprezentată de către modelarea fonetică. Două foneme aparent identice vor trebui să fie tratate separat, deoarece pronunția lor diferă in funcție de limbă. Cea de-a doua abordare presupune crearea unor modele individuale, specifice fiecărei limbi. Însă, premergător etapei de transcriere, va fi necesară identificarea limbii vorbite pentru a stabili modelul acustic ce va fi folosit.

4.9 Sumar al realizărilor / rezultatelor

Nr.	Denumirea rezultatului	Denumire activitate	Procent de realizare
1.	*D3.1 - Raport despre starea artei in recunoasterea multilingva a vorbirii*	*Activitatea 3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti*	*100%*
2.	*D3.10 - Articole stiintifice in conferinte si jurnale*	*Activitatea 3.6. Diseminarea rezultatelor proiectului (partea I)*	*100%*
3.	*D3.11 - Corpus de vorbire anotat si aliniat pentru limba romana (partea a II-a)*	*Activitatea 3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)* *Activitatea 3.2 Validarea manuala a corpusului de vorbire*	*100%*
4.	*D3.12 - Model functional pentru controlerul de voce*	*Activitatea 3.3. Implementarea unui controller de voce (partea a II-a)*	*100%*
5.	*D3.13 - Model functional de recunoastere a vorbirii la distanta, robust la zgomot*	*Activitatea 3.4. Implementarea parametrilor PNCC si a tehnicilor de reducere a zgomotului (partea a II-a)* *Activitatea 3.5. Compararea si evaluarea performantelor tehnicilor implementate*	*100%*

4.10 Bibliografie

[Boros, 2013] Boroş, Tiberiu and Dumitrescu, Ştefan Daniel. Improving the RACAI Neural Network MSD Tagger. In Engineering Applications of Neural Networks (Lazaros Iliadis and Harris Papadopoulos and Chrisina Jayne). Springer, vol. 383, pp. 42--51, 2013.

[Rousseau, 2012] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus", in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), May 2012.

[Vacher, 2015] Michel Vacher, et.al. „Speech and Speaker Recognition for Home Automation: Preliminary Results," in Proceedings of The 8th InternationalConference Speech Technology and Human-Computer Dialogue "SpeD 2015", Oct 2015.

[Williams, 2015] W. Williams, N. Prasad, D. Mrva, T. Ash, and T. Robinson, "Scaling recurrent neural network language models," arXiv preprint arXiv:1502.00512, 2015.

5 Documente in extenso

Nr.	Denumire anexă	Activitate asociata	Tip*
1.	*50.7.1. Raport despre starea artei in recunoasterea multilingva a vorbirii v4.docx*	*Activitatea 3.7. Studiul algoritmilor state-of-the-art in detectia multilingva a termenilor vorbiti*	*Raport cercetare*
2.	*50.7.2. Descriere corpus vorbire si sistem TTS.docx*	*Activitatea 3.1 Segmentarea si adnotarea corpusului de vorbire (partea a II-a)* *Activitatea 3.2 Validarea manuala a corpusului de vorbire*	*Descriere livrabil*
3.	*Materiale multimedia ce demonstreaza modelul functional pentru controlerul de voce (livrabil D3.12)*	*Activitatea 3.3. Implementarea unui controller de voce (partea a II-a)*	*Filme, poze*

6 Anexa 1 – Gramatica FSG Cassandra

```
#JSGF V1.0;

/**
 * JSGF Grammar for the FSG ANVSIB demo
 */

grammar functii;
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<exterior> = <irigatie> | <statie_meteo> | <usa_principala> | <usa_garaj> | <poarta_principala> | <moduri>;

<irigatie> = (porneşte|opreşte|activează|dezactivează) (stropitorile|irigaţia|sistemul de irigaţie) | dă drumul (la stropitori | stropitorilor|la irigaţie|irigaţiei|la sistemul de irigaţie|sistemului de irigaţie);
<statie_meteo> = (cât|care|ce) ([este] umiditatea ((în|din) (casă|cameră)|[de] afară)| [este] radiaţia solară | radiaţie solară [este]);
<usa_principala> = (încuie|descuie|închide|deschide|cameră) (uşa de la intrare|intrarea|[pe] uşa principală);
<usa_garaj> = (încuie|descuie|închide|deschide) (uşa de la garaj|garajul);
<poarta_principala> = (încuie|descuie|închide|deschide) (poarta de la intrare|poarta principală);
<moduri> = (porneşte|activează) modul de (citit|cină|film);
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<securitate> = <centrala_efractie> | <cctv> | <simulare_prezenta> | <salvare> | <poliţie> | <pompieri>;

<centrala_efractie> = (armează|dezarmează|activează|dezactivează) partiţia [[cu] numărul] <numar_cifra> | (care este statusul partiţiei|ce status are partiţia|cum este partiţia|cum e partiţia) <numar_cifra> | (status partiţie|statusul partiţiei) <numar_cifra>;
<cctv> = (afişează|arată|arată-mi|pune|(zoom-in|zoom-out) pe|[fă] (zoom-in|zoom-out) pe|depărtează|apropie|măreşte|micşorează) (camera <numar_cifra>|toate camerele|(uşa|poarta) (principală|de la intrare)) [pe (televizor|intercom)];
<simulare_prezenta> = (porneşte|opreşte|activează|dezactivează) (simularea de prezenţă|simulatorul de prezenţă|simularea prezenţei|simulatorul prezenţei) | dă drumul (la simularea de prezenţă|la simulatorul de prezenţă|la simularea prezenţei|simulării prezenţei);
<salvare> = (sună | cheamă ) ([la] salvare | salvarea);
<poliţie> = (sună | cheamă ) ([la] poliţie | poliţia);
<pompieri> = (sună | cheamă ) (pompierii | [la] pompieri);
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<multimedia> = <sursa> | <lista_redare> | <volum>;

<sursa> = (schimbă|comută) sursa [televizorului] pe (hdmi|vga|plug-in);
<lista_redare> = rulează lista de redare;
<volum> = (măreşte|creşte|micşorează|scade|dă mai (tare|încet)) (volumul|sonorul) [televizorul|televizorului|muzicii|la televizor|la muzica] | (dă mai (tare|încet)|porneşte|opreşte) (televizorul|muzica) | [dă] [televizorul|muzica|volumul|sonorul] mai (tare|încet);
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<hvac> = <aer_conditionat> | <viteza_ventilator> | <blinds> | <ferestre> | <contor_termic>;

<viteza_ventilator> = (setează|schimbă) (viteza|temperatura) (ventilatorului|camerei) la (<numar_cifra>|<numar_zeci>|<numar_ac>) (la sută|de grade);
<aer_conditionat> = (setează|dă|porneşte|opreşte|schimbă|pune) aerul condiţionat [(pe|la) <numar_ac> [de] grade|mai cald|mai rece|pe [modul] [de] (răcire|încălzire|ventilaţie)];
<blinds> = (închide|deschide) ((draperiile|obloanele|jaluzelele) [pe jumătate]) | (înclină|apleacă) (jaluzelele|obloanele) la <numar_zeci> la sută | trage (draperiile|jaluzelele);
<ferestre> = (închide|deschide) (fereastra|[toate] ferestrele|geamul|[toate] geamurile);
<contor_termic> = care este temperatura camerei|câte grade sunt [afară|în cameră|aici]|(e|este) (frig|cald) afară|citeşte temperatura camerei;
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<electric> = <lumini> | <culoare_lumina> | <dimmer> | <alimentare> | <panouri_solare>;

<lumini> = (aprinde|stinge|opreşte|porneşte|închide|deschide) (lumina|[toate] luminile) [[de] la parter|din casă];
<culoare_lumina> = (schimbă|aprinde|porneşte) (culoarea luminii|lumina) (roşie|în roşu|[în] verde|[în] albastru|[în] albastră|[la] normală|la normal);
<dimmer> = (măreşte|creşte|micşorează|scade) luminozitatea [până] la <numar_zeci> la sută;
<alimentare> = (porneşte|reia|opreşte) (alimentarea|încărcarea bateriei);
<panouri_solare> = care este starea bateriei|este încărcată bateria;
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
<numar_cifra> = zero | unu | una | doi | două | trei | patru | cinci | şase | şapte | opt | nouă;
<numar_zeci> = zece | douăzeci | treizeci | patruzeci | cincizeci | şaizeci | şaptezeci | optzeci | nouăzeci | [o] sută;
<numar_ac> = zece | unsprezece | doisprezece | douăsprezece | treisprezece | paisprezece | cincisprezece | şaisprezece | şaptesprezece | optsprezece | nouăsprezece | douăzeci | douăzeci_şi_unu | douăzeci_şi_doi | douăzeci_şi_trei | douăzeci_şi_patru | douăzeci_şi_cinci | douăzeci_şi_şase | douăzeci_şi_şapte | douăzeci_şi_opt | douăzeci_şi_nouă | treizeci | treizeci_şi_unu | treizeci_şi_doi | treizeci_şi_trei | treizeci_şi_patru | treizeci_şi_cinci | treizeci_şi_şase | treizeci_şi_şapte | treizeci_şi_opt | treizeci_şi_nouă | patruzeci;
//-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
public <comandaCasa> = casandra (<exterior> | <securitate> | <multimedia> | <hvac> | <electric>);
```