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: *02.12.2014*
Etapa : *1/2014*
Activitatea / activităţile:

- *Activitatea 1.1. Cerinţe arhitecturale*
- *Activitatea 1.2. Modele experimentale pentru componentele sistemului*
- *Activitatea 1.3. Proiectarea corpusului de vorbire (partea I)*
- *Activitatea 1.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteză şi recunoaştere a vorbirii (partea I)*
- *Activitatea 1.5. Achiziţionarea unei baze de date de vorbire în condiţii reale (partea I)*
- *Activitatea 1.6. Diseminarea rezultatelor proiectului (partea I)*

Număr contract: *32 / 2014*
Acord de colaborare nr. 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
Coordonator proiect: *Universitatea Politehnica Bucureşti*

Director de Proiect: *Horia Cucu*

Cuprins

1 Obiectivele activităților desfăşurate

- *Activitatea 1.1: Definirea cerințelor arhitecturale ale sistemului*
- *Activitatea 1.2: Crearea modelor experimentale pentru componentele sistemului*
- *Activitatea 1.3: Proiectarea unui corpus de vorbire în scopul sintezei vorbirii*
- *Activitatea 1.4: Crearea unui proxy de voce pentru integrarea VoIP a sistemelor de sinteză şi recunoaştere a vorbirii*
- *Activitatea 1.5: Achizitionarea unei baze de date de vorbire in conditii reale în scopul recunoaşterii de vorbire*
- *Activitatea 1.6: Diseminarea rezultatelor proiectului*

2 Perioada de desfăşurare şi personalul implicat

Nr.	Denumire activitate	Nume si prenume	Perioada
1.	*Activitatea 1.1. Cerințe arhitecturale*	*CO:Cucu Horia; Buzo Andi; Marinescu Radu; Dogariu Mihai; Minca Florentina* *P1:* Marian Bădulescu; Marius Petrescu *P2: Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.07.2014 – 31.12.2014*
2.	*Activitatea 1.2. Modele experimentale pentru componentele sistemului*	*CO:Cucu Horia; Buzo Andi; Marinescu Radu; Dogariu Mihai; Minca Florentina* *P1:* Marian Bădulescu; Marius Petrescu *P2: Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.07.2014 – 31.12.2014*
3.	*Activitatea 1.3. Proiectarea corpusului de vorbire (partea I)*	*Ştefan Daniel Dumitrescu; Tiberiu Boroş; Dan Tufiş*	*01.07.2014 – 31.12.2014*
4.	*Activitatea 1.4. Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteză şi recunoaştere a vorbirii (partea I)*	*Marian Bădulescu; Marius Petrescu*	*01.07.2014 – 31.12.2014*
5.	*Activitatea 1.5. Achiziționarea unei baze de date de vorbire in condiții reale (partea I)*	*CO:Cucu Horia; Buzo Andi; Marinescu Radu; Dogariu Mihai; Minca Florentina*	*01.07.2014 – 31.12.2014*
6.	*Activitatea 1.6. Diseminarea rezultatelor proiectului (partea I)*	*CO:Cucu Horia; Buzo Andi; Marinescu Radu; Dogariu Mihai; Minca Florentina*	*01.07.2014 – 31.12.2014*

3 Rezumat activități

3.1 Cerințe arhitecturale (Activitatea 1.1.)

În această activitate, partenerii au definit şi au căzut de acord asupra arhitecturii sistemului. S-a decis independența între componentele dezvoltate de parteneri diferiți cu scopul de a uşura dezvoltarea lor. În consecință s-a stabilit modul de interfațare între componente prin protocoale standard. S-au stabilit scenariile de utilizare şi fluxul de date. În raport se menționează, de asemenea, criteriile după care au fost luate unele decizii la nivel de arhitectură şi la nivel de terminale. Criteriile au fost de natură comercială, de interoperabilitate, de flexibilitate şi de fiabilitate.

Activitatea a fost realizată în proporție de 100%.

3.2 Modele experimentale pentru componentele sistemului (Activitatea 1.2.)

În cadrul Activității 1.2. (Modele experimentale pentru componentele sistemului) a fost creat un sistem automat de recunoaştere de vorbire (Automatic Speech Recognition - ASR) şi s-a evaluat performanța acestuia pe o bază de date discutată mai pe larg în secțiunea despre Activitatea 1.5. Problema adresată se referă la crearea unui sistem ASR pentru vorbire continuă, independentă de vorbitor în limba română care să funcționeze în interiorul unei case inteligente şi să recunoască un set de comenzi specifice sistemelor automatizate din interiorul casei. În acest fel se doreşte punerea sub formă textuală a comenzilor vocale pe care un utilizator le dă casei sale. Având la dispoziție resursele necesare antrenării unui astfel de sistem ASR am adăugat şi o alternativă la modelul de limbă utilizat, reprezentată de gramatica cu stări finite.

Activitatea 1.2 a cuprins următoarele subactivități:

- crearea unui sistem ASR de vorbire continuă, independentă de vorbitor pentru limba română;
  - crearea unui model acustic, model de limbă cu n-grame şi un dicționar fonetic;
  - crearea unei gramatici cu stări finite;
  - introducerea unui cuvânt cheie în frazele ce conțin comenzi pentru casă pentru a elimina din discuție toate frazele ce nu au legătură cu comanda sistemului;
- evaluarea sistemului;
  - stabilirea unei metrici de evaluare adecvată pentru scopul proiectului;
  - rularea testelor cu diferite configurații şi stabilirea configurației optime;
  - compararea rezultatelor obținute folosind modelul de limbă cu n-grame cu cele obținute folosind gramatica cu stări finite.

Modelul experimental de sinteză a vorbirii este bazat pe sistemul RACAI existent. Astfel, initial prezentăm sistemul existent (descriind ambele componente: modulul de preprocesare text precum şi modulul de sinteză efectivă a vocii). În ceea ce priveşte componenta de sinteză a vorbirii, în cadrul acestei activități, am extins suportul pentru sinteză parametrică statistică (obiectiv tehnic), investigarea unor metode de îmbunătățire a calității prozodice a vocii (obiectiv ştiințific) şi, în final, dezvoltarea unui backend hibrid (ce combină unități naturale de voce cu unități sintetice) care să asigure o calitate superioară a vocii obținute (naturalețe), respectând în acelaşi timp cerințele prozodice (lucru ce nu putea fi realizat cu de un sistem concatenativ datorită limitărilor impuse de dimensiunea corpusului acustic). Am ales utilizarea back-endului HMM Speech Synthesis System cunoscut şi sub numele HMM Tripple ‚S' (HTS). In cadrul prezentării extinse sunt detaliați parametrii folositi pentru antrenarea sistemului. In finalul prezentării sunt expuse corpusurile proprii si externe deja existente necesare pentru antrenare. Modelul experimental astfel obținut este disponibil online.

Controllerul de voce este partea inteligentă a sistemului. În această etapă a fost construit un model experimental care are funcționalitatea descrisă mai jos. Controlerul de voce funcționează de inițiativă proprie la pornirea sistemului aplicând configurația implicită (default). În cursul funcționării normale, el are următoarele sarcini de realizat:

- să primească prin VoIP fluxul audio de la microfoanele instalate în casa inteligentă;
- să dirijeze fluxul audio către serverul ASR;
- să primească raspuns de la serverul ASR;

- să convertească răspunsul serverului ASR într-o comandă pentru sistemul casei automate (Building Operation System);
- să primească confirmare sau răspuns de la sistemul casei automate;
- să deducă răspunsul destinat utilizatorului din răspunsul sistemului casei automate;
- să trimită răspunsul text către serverul TTS;
- să primească fluxul audio de la sistemul TTS;
- să dirijeze fluxul audio primit de la sistemul TTS către difuzoare prin VoIP.

Activitatea a fost realizată în proporție de 100%.

3.3 Proiectarea corpusului de vorbire (partea I) (Activitatea 1.3.)

A doua activitate principală coordonată de RACAI este proiectarea corpusului de vorbire. Deoarece sistemele de sinteză parametrice sau concatenative folosesc corpusuri pentru antrenare, dimensiunea şi calitatea acestor corpusuri sunt esențiale in rezultatele acustice ale acestor sisteme. Astfel, activitatea de proiectare este un pas esențial pentru crearea unui sistem de sinteză de performanță. În cadrul descrierii sunt prezentate caracteristicile principale ale unui astfel de corpus. Este discutată distribuția datelor de intrare, acoperirea unui astfel de corpus din punct de vedere al domeniilor lingvistice, din punct de vedere al acoperirii inventarului fonetic al limbii române precum si contextul necesar fiecărei unități fonetice. Este prezentat procesul de selecție al datelor pentru pregătirea corpusului.

Activitatea a fost realizată în proporție de 100%.

3.4 Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteză şi recunoaştere a vorbirii (partea I) (Activitatea 1.4.)

Proxy-ul de voce are rolul de a fi interfața între utilizatori şi cele 2 sisteme de voce (ASR sau TTS) aşa cum este arătat în Fig. 1. Acest modul a fost introdus pentru a respecta principiul de interoperabilitate şi independență între sisteme. S-a folosit o tehnologie solidă de VoIP pentru vehicularea fluxului audio. Această tehnologie permite unificarea modului de comunicație între difuzoare şi controllerul de voce şi între telefoane şi controllerul de voce.

Activitatea a fost realizată în proporție de 100%.

3.5 Achiziționarea unei baze de date de vorbire in condiții reale (partea I) (Activitatea 1.5.)

În cadrul activității 1.5 (Achiziționarea unei baze de date de vorbire in condiții reale (partea I)) a fost creată o bază de date de înregistrări audio de propoziții în condiții de laborator. Frazele înregistrate reprezintă un set de comenzi pe care utilizatorul le poate da casei inteligente. Înregistrările în condiții de laborator se justifică deoarece ele sunt folosite pentru a testa sistemul ASR în depistarea şi soluționarea eventualelor erori de proiectare. Contextul frazelor folosite în această etapă este similar cu cel al frazelor care vor fi folosite în condiții reale, de către utilizator.

Activitatea 1.5 a cuprins următoarele subactivități:

- întocmirea unei liste cu toate echipamentele din casă ce pot fi automatizate
- stabilirea acțiunilor ce pot fi desfăşurate 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
- selectarea studenților în vederea realizării înregistrărilor audio
- înregistrarea propriu zisă a clipurilor audio (utilizând o aplicație online de înregistrări audio realizată anterior de UPB)

Activitatea a fost realizată în proporție de 100%.

3.6 Diseminarea rezultatelor proiectului (partea I) (Activitatea 1.6.)

În cadrul acestei activități a fost creată pagina web a proiectului care conține informații cheie cum ar fi:

- finanțatorul proiectului;
- descrierea partenerilor şi rolul lor în proiect;
- rezumatul proiectului;
- durata proiectului.

Proiectul a fost făcut cunoscut în forumurile în care laboratorul SpeeD are acces.

Adresa paginii web este: **http://speed.pub.ro/anvsib**

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

4.1 Cerințe arhitecturale (Activitatea 1.1.)

Arhitectura sistemului este descrisă în Fig. 1. Sistemul are următoarele componente principale: sistemul de recunoaştere a vorbirii (ASR), sistemul de sinteză a vorbirii (TTS), controller-ul de voce, proxy-ul de voce, serverul SIP (Session Initiation Protocol) şi terminalele de input şi output. Comunicația între echipamente se face utilizând tehnologii VoIP (Voice over IP).

Utilizatorul poate da comenzi vocale care sunt captate de microfoane care sunt instalate în camera inteligentă. De asemenea, comenzile se pot da şi din afara casei printr-un telefon mobil sunând la un număr predefinit. Toate aceste terminale sunt conectate la acelaşi server SIP. Semnalul vocal este apoi transferat controlerului de voce prin intermediul proxy-ului de voce. Controllerul decide ca vocea să fie trimisă către componenta ASR spre recunoaştere. Componenta ASR răspunde cu transcrierea vorbirii din semnalul vocal. Această transcriere reprezintă comanda dată de utilizator sau un mesaj care să reprezinte lipsa de comandă. Dacă semnalul vocal conține o comandă, atunci controllerul de voce trimite aceste comandă către BOS (interfața cu echipamentele comandabile prin KNX). Aplicarea comenzii este însoțită de o confirmare care ajunge în controllerul de voce. Dacă comanda executată este prevăzută şi cu răspuns vocal, atunci controllerul de voce trimite mesajul de pronunțat către sistemul TTS care răspunde la rândul lui cu un semnal vocal care conține mesajul. Acest mesaj este trimis către difuzoare, care sunt instalate în camera inteligentă, prin intermediul proxy-ului de voce şi serverul SIP.

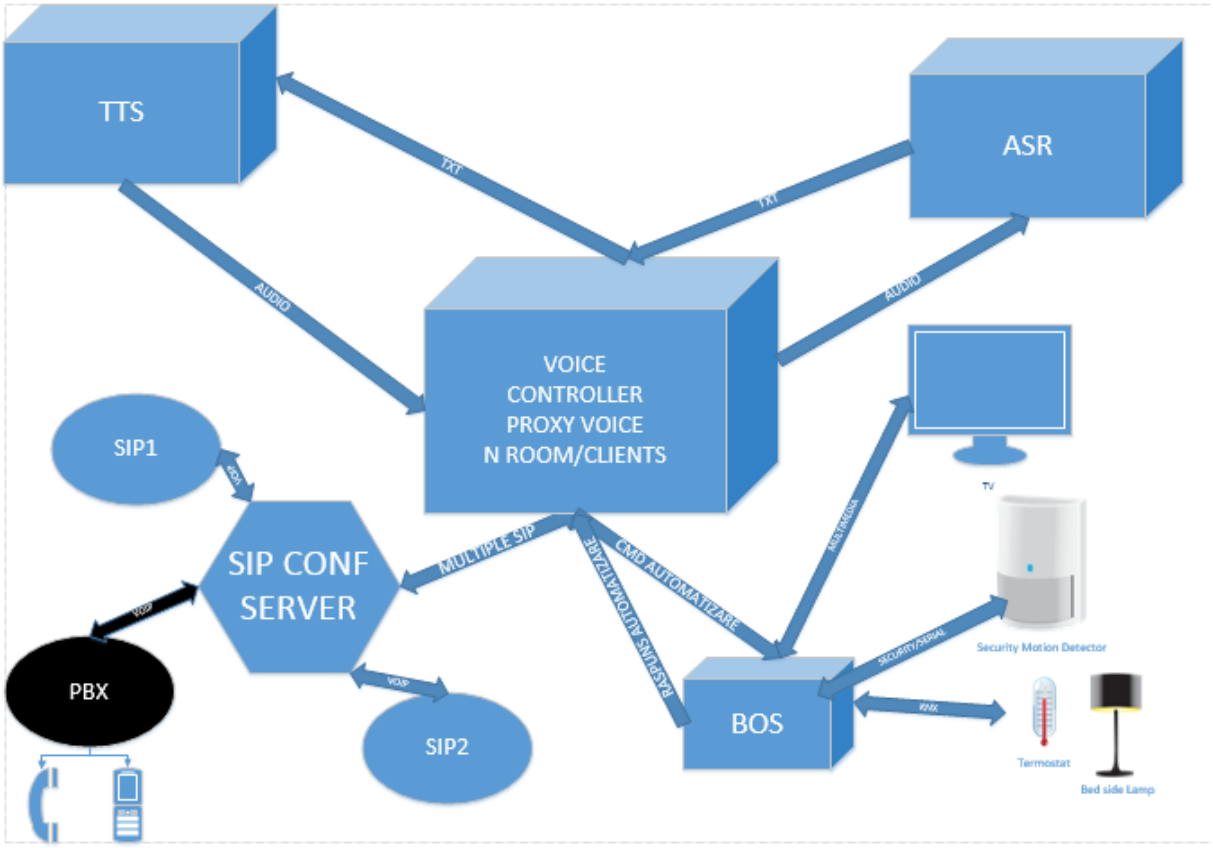

Figura 1. Arhitectura generală a sistemului

Cerințele impuse asupra ASR şi TTS sunt legate de timpul de răspuns (în timp real), parametrii semnalului vocal (16 KHz, 16 biți/eşantion) şi interfață (folosind interfețe standard). Cerințele asupra restului de echipamente au fost definite pe baza următoarelor criterii:

1. Criterii comerciale
2. Criterii de interoperabilitate
3. Criterii de fiabilitate/flexibilitate

Sistemul trebuie să respecte urmatoarele *criterii comerciale*:

- Echipamentele folosite să fie uşor de procurat şi la prețuri neprohibitive.
- Echipamentele utilizate să asigure mobilitatea utilizatorului şi principiul mâinilor libere.
- Controllerul de voce trebuie sa fie capabil să se conecteze şi în exteriorul casei pentru a prezenta diverse alarme (inundație, incendiu, efracție, etc.).
- Controllerul de voce trebuie să accepte din exterior comenzi (de exemplu: armează alarama / simulează prezența).

Aceste criterii au avut următorul impact asupra definirea sistemului:

- Am luat decizia de a testa pe echipamente cu suport android (telefoane mobile cu căşti bluetooth sau tablete).
- Se vor utiliza tehnologii existente pentru care există deja support cum ar fi SIP (pentru semnalizare) şi alte protocoale standard.
- Echipamentele sunt instalate în pereți sau tavan ca să nu necesite utilizarea mainilor.
- Utilizând un server de VOIP conectat la un PBX clasic acesta poate trimite alarme cu ajutorul unui simplu card.
- Utilizând un card de conectare la PBX, centrala VOIP poate accepta un apel din exterior pentru a putea permite recunoaşterea şi sintetizarea prin această conexiune.

Criteriile de interoperabilitate sunt:

- Să existe o comunicație standardizată între sistemul ASR şi TTS atât pentru comenzi cât şi pentru fluxurile audio.
- Să existe posibilitatea conectării cu terminale vandabile (costuri rezonabile).
- Să existe posibilitatea interconectării sistemului cu terminale clasice (telefonie analogica, mobila, etc.).

Aceste criterii au avut următorul impact asupra definirea sistemului:

- Deşi componentele pot rula pe acelaşi calculator, ele sunt entități software independente. Astfel sistemul devine flexibil şi permite re-împrospătarea fiecărei componente independent.
- Am luat decizia utilizării standardului MRCP şi SIP pentru transmiterea şi recepționarea textului şi a semnalului audio.
- Atât telefoanele mobile cât şi microfoanele instalate în casă pot fi folosite pentru a da comenzi de automatizare prin intermediul vocii umane sau se poate obține statusul sistemelor de automatizare (temperatura din camera, armarea partitiilor, inundație, incendiu, etc.).
- un simplu telefon fix sau mobil poate fi folosit ca şi terminal de intrare prin intermediul cardului PBX din centrala VOIP.

Pentru *flexibilitatea si fiabilitatea* sistemului am tinut cont de urmatoarele aspecte

- Sistemul trebuie să poata funcționa 24/365 fără a necesita intervenție umană.
- Configurarea sistemului să fie uşor de făcut atât pentru o clădire cu 1 canal de recunoaştere voce sau sinteza cât şi pentru un proiect cu 100 de canale de comunicație.
- Sistemul sa poata fi configurat parametrizabil şi adaptiv.

Aceste criterii au avut următorul impact asupra definirea sistemului:

- Utilizarea unor echipamente deja testate pentru condiții de 24/365 de la producători consacrați.
- Utilizând soluția ASTERISK nu avem limita la numărul de interioare sau terminale folosite. Totul depinde de configurația hardware a sistemului pe care este instalat.
- Avand în vedere că fiecare utilizator doreşte să folosească un alt tip de limbaj am luat decizia utilizării unei interfețe web pentru configurare şi personalizare de fiecare utilizator.

4.2 Modele experimentale pentru componentele sistemului (Activitatea 1.2.)

4.2.1 Modelul experimental ASR

Scopul principal al unui sistem ASR este de a reda sub formă de text ceea ce a fost rostit într-o secvență audio sau, altfel spus, de a translata vorbirea în text (Speech-to-Text). Acest modul primeşte ca semnal de intrare o înregistrare audio sau un flux audio în timp real, după caz, şi redă la ieşire transcrierea textuală a mesajului ce a fost comunicat sub formă de vorbire.

Realizarea unui astfel de sistem ridică cu atât mai multe dificultăți cu cât condițiile în care se realizează testarea sa se îndepărtează mai mult de condițiile în care a fost antrenat. Totuşi, principiul de funcționare care stă la baza unui astfel sistem de recunoaştere de vorbire rămâne acelaşi, iar acest cadru fundamental poate fi testat şi în condiții de laborator pentru a-i observa performanțele, urmând ca în faza următoare să se treacă la testarea în condiții reale.

Nume model acustic	branch1	mixModels	spk01-20
Durată totală [ore]	87	73	40
Nr. vorbitori	200+	70	17
Tip vorbire	dictare + vorbire spontană	dictare	dictare
Nr. senone	2000	4000	4000
Nr. mixturi gaussiene	16	64	32
Nr. stări/HMM	3	3	3

Prima fază a sistemului ASR proiectat are ca scop verificarea funcționării framework-ului şi depistarea defectelor ce afectează principiul de funcționare. Se doreşte stabilirea unui domeniu al parametrilor în care sistemul urmează să funcționeze şi adaptarea acestuia astfel încât să prezinte o robustețe cât mai mare la variațiunile ce pot apărea în cazurile reale.

În vederea realizării sistemului ASR s-au folosit următoarele:

- trei modele acustice cu următoarele specificații:
  - cele 3 modele acustice sunt bazate pe HMM-uri şi au fost antrenate cu fraze cu contexte diferite (politică, literatură, buletin meteo etc).
- un dicționar fonetic cu 15,000 cuvinte în faza testării cu modelul de limbă
- un model de limbă cu 64,000 unigrame, 13,000,000 bigrame, 16,000,000trigrame obținut prin interpolarea ponderata a modelului de limbă format exclusiv din comenzi (pondere: 90%) cu un model de limbă mai general, ce conține 64,000 cuvinte diferite (pondere: 10%).
- ca soluție alternativă pentru modelul de limbă din faza de testare am propus utilizarea unei gramatici cu stări finite şi, implicit, a unui nou dicționar fonetic cu 900 cuvinte, format pe baza listei de comenzi.

Urmărind destinația concretă în care va fi folosit sistemul de recunoaştere, apar anumite aspecte ce trebuie menționate, întrucât ele contribuie la realizarea întregului framework şi a condițiilor de testare:

1. sesizarea momentului în care utilizatorul i se adresează casei
2. modul de achiziție al informației şi transportul ei până la ASR
3. contextul în care vor fi plasate semnalele vocale ce se aşteaptă la intrarea sistemului

Fiecare dintre aceste aspecte va fi discutat în parte în secțiunea următoare.

1. Un prim aspect de o importanță semnificativă este mediul înconjurător din care se va extrage semnalul vocal folosit pentru realizarea recunoaşterii şi anume incinta casei inteligente. Introducerea unui sistem ASR în interiorul unei case în care diferite persoane urmează să îşi desfăşoare activitățile ridică o problemă cu privire la sesizarea acelui interval de vorbire din discursul unei persoane în care aceasta i se adresează casei. Acest lucru este dorit pentru a nu interpreta şi recunoaşterea de vorbire ce a avut loc pe fraze care nu au legătură cu scopul proiectului drept comenzi către casă şi a utiliza, în acest mod, resurse în mod inutil. Mai mult, această implementare scade şi posibilitatea de a recunoaşte, în mod eronat, comenzi care sunt, de fapt, doar secvențe din conversația a două sau mai multe persoane (numărul de *false alarms*).

Pentru a evita aceste aspecte am adăugat un cuvânt cheie la nivelul fiecărei fraze care va fi considerată comandă pentru casa inteligentă. Alegerea acestui cuvânt a avut în vedere selecția unui nume propriu, rar întâlnit în limba română pentru a nu se confunda cu adresarea către o persoană ce poate fi în acel moment în casă şi poartă acelaşi nume ca cel al casei, uşor de reținut de către utilizator şi care nu introduce o varietate mare în modul în care poate fi pronunțat. Astfel, toate comenzile vor fi de forma:

Casandra, #comandă#!

#comandă# va fi înlocuit cu o comandă propriu-zisă către casă. Se observă, deci, că numele ales pentru adresarea către casă este “Casandra”, iar semnalul vocal ce va fi recunoscut de către sistemul ASR va avea, obligatoriu, acest cuvânt în componență.

Toate comenzile vor începe cu cuvântul cheie ‘Casandra’ pentru a anunța sistemul că utilizatorul i se adresează, în mod direct, în felul acesta eliminând o mare parte din confuziile care pot apărea în legătură cu

vorbirea cotidiană. Folosind un nume propriu se obține și un grad de familiaritate mai mare pentru a ajuta utilizatorul să fie cât mai natural în vorbire atunci când i se adresează casei.

2. Referitor la modul de achiziție al informației și cum va fi aceasta transportată până la ASR precizăm că între sistemul de achiziție a datelor și cel de prelucrare a lor este o legătură de tip client-server, unde casa inteligentă este clientul, iar mașina pe care se va găsi ASR-ul va avea rolul de server. În prima fază de proiectare, sistemul ASR va face recunoaștere continuă asupra fluxului audio pe care îl primește în timp real de la sistemul de achiziție.

La nivelul casei se va afla sistemul Speech Controller, care va gestiona fluxurile audio de la echipamentele de achiziție din casă către server-ul pe care se află sistemul ASR și tot el va fi destinatarul răspunsului trimis de către ASR. În urma recunoașterii de vorbire se va obține, sub formă scrisă sau ca identificator de comandă, comanda pe care utilizatorul a dat-o casei, iar această informație va fi trimisă de la server-ul care se ocupă cu recunoașterea vorbirii înapoi către Speech Controller.

S-a recurs la acest mod de lucru (achiziție de date local și prelucrarea lor remote) pentru a nu obliga utilizatorul să adauge echipamente suplimentare în interiorul casei. De asemenea, această soluție permite ca un singur echipament (server) să deservească mai multe case inteligente (clienți) fără o creștere a costurilor de implementare în ceea ce privește adăugarea unui nou sistem ASR.

3. Contextul în care vor fi plasate semnalele vocale ce se așteaptă la intrarea ASR-ului este unul limitat de către obiectivul final al sistemului de recunoaștere și anume de a recunoaște comenzi pe care un utilizator le poate da casei inteligente, adresându-se cu privire la un set finit de echipamente. Comanda, în sine, va fi orientată pe sistemele care vor fi automatizate, urmând un model cât mai apropiat de vorbirea naturală, de exemplu: “Casandra, închide geamul!”, “Casandra, aprinde lumina!”, “Casandra, pornește televizorul!”, etc.

Pentru deservirea unei comenzi au fost luate în calcul diferite formulări posibile, întrucât pot exista mai multe metode prin care se poate cere un același lucru de la casă. De asemenea, s-a ținut cont și de faptul că vorbirea colocvială introduce mai multe variante de comandă asupra unui același echipament (“Casandra, pornește stropitorile!” vs. “Casandra, dă drumul la stropitori!”). Lista cu comenzile posibile pe care le va primi casa inteligentă a fost făcută pe baza listei de echipamente ce pot fi configurate de către aceasta. Din lista cu comenzi posibile a fost derivată și gramatica cu stări finite care a fost folosită ca alternativă la modelul de limbă cu n-grame.

Tabelul I. Rezultatele obținute pentru diferite configurații ale sistemului ASR

Baza de date testata	Modelul acustic folosit	FSG/LM	Miss probability (%)	False alarm probability (%)	Correctly recognized (%)	DER (%)
homeAuto2	branch1	FSG	29.1	0	60.9	47.78
homeAuto2	mixModels	FSG	28.5	0	71.5	39.86
homeAuto2	spk01-20	FSG	41.7	0	58.3	71.52
homeAuto2	branch1	LM	59.5	0	40.5	146.91
homeAuto2	mixModels	LM	62.9	0	37.1	169.54
homeAuto2	spk01-20	LM	78.4	0	21.6	362.96
databaseOmega	branch1	FSG	0	0	100	0
databaseOmega	mixModels	FSG	0	0	100	0
databaseOmega	spk01-20	FSG	0	0.04	99.96	0.04
databaseOmega	branch1	LM	0	0	100	0
databaseOmega	mixModels	LM	0	0	100	0
databaseOmega	spk01-20	LM	0	0	100	0

Starea actuală a sistemului de recunoaştere de vorbire este completă, dar ea abordează întregul proiect din perspectiva unui proiect pilot din care urmează să selectăm cele mai bune configuraţii, să identificăm factorii ce pot veni în detrimentul proiectului, în ansamblu şi să oferim soluţii pentru problemele apărute. Ulterior, proiectul pilot va fi extins până la nivelul în care va fi gata de a aborda condiţiile reale întâlnite într-o casă inteligentă şi testat în mediul în care va fi livrat clienţilor.

Sumarizând rezultatele obţinute în urma testării sistemului ASR creat cu cele 3 modele acustice am întocmit Tabelul I (DER ideal este 0).

4.2.2 Modelul experimental TTS

Înainte de a prezenta contribuţiile tehnice şi ştiinţifice aduse în cadrul proiectului vom face o scurtă recapitulare a elementelor existente înainte de începerea derulării ANVSIB. Aşa cum a fost prezentat anterior, înainte de începerea proiectului ANVSIB, RACAI dispunea de un sistem de sinteză a vorbirii pornind de la text format din două componente principale: un sistem de pre-procesare a textului (frontend) şi un sistem de sinteză a vorbirii (backend). Sistemul de pre-procesare a textului este construit pe un clasificator de tip perceptron având suport pentru algoritmul de antrenare MIRA (Margin Infused Relaxed Algorithm). Fiecare pas implicat în procesul de pregătire a textului de intrare pentru sinteza efectivă este realizat folosit o strategie specifică de etichetare, precum codificarea onset-nucleus-coda (ONC) pentru silabaţie (Bartlett et al., 2008) (99% pentru limba română la cuvintele OOV), alinieri EM cu etichetare secvenţială pentru conversia L2S (Jiampojamarn et al., 2009) (96.29% pentru limba română la cuvintele în afara vocabularului) şi, de asemenea, abordări originale ale lematizării (92.83%) şi ale predicţiei accentului lexical (Boroş et al., 2013) (98.80%). Analizorul morfo-sintactic implementat de sistemul RACAI foloseşte o metodă originală prezentată în (Boroş et al., 2013) pentru etichetarea cu un set mare de etichete morfosintactice, bazată pe reţele neuronale, ceea ce facilitează compatibilitatea cu limbile cu morfologie bogată. Sistemul de sinteză a vorbirii folosea o implementare de sinteză concatenativă pe baza algoritmului PSOLA ce folosea unităţi de mostre de voce obţinute prin alinierea automată a corpusului de intrare. Această aliniere era realizată folosind instrumentul Hidden Markov Model Toolkit (HTK).

În ceea ce priveşte componenta de sinteză a vorbirii, RACAI şi-a propus extinderea suportului pentru sinteză parametrică statistică (obiectiv tehnic), investigarea unor metode de îmbunătăţire a calităţii prozodice a vocii (obiectiv ştiinţific) şi, în final, dezvoltarea unui backend hibrid (ce combină unităţi naturale de voce cu unităţi sintetice) care să asigure o calitate superioară a vocii obţinute (naturaleţe), respectând în acelaşi timp cerinţele prozodice (lucru ce nu putea fi realizat cu de un sistem concatenativ datorită limitărilor impuse de dimensiunea corpusului acustic).

În sinteza parametrică statistică, semnalul acustic este reprezentat folosind parametrii spectrali utilizaţi şi în recunoaşterea vorbirii, a frecvenţei fundamentale a vocii şi a duratei explicite (nu dedusă implicit prin folosirea numărului de stări în care se staţionează) pentru fiecare fonem în parte (King, 2010). Antrenarea unui astfel de sistem constă în realizarea unui model statistic în care parametri vocali (media şi varianţa) sunt modelaţi în timp pe baza unui set de observaţii, care în acest caz constau în valori multinomiale extrase automat din text (de către frontend). Pentru generarea semnalului acustic, un decodor este folosit să aleagă o secvenţă de stări (parametri acustici) pe baza unui set de trăsături, parametri acustici fiind ulterior convertiţi în voce folosind o tehnică specifică vocoderelor. Printre cele mai cunoscute metode folosite la inversarea acestor parametri numărăm Mel-Log Spectral Aproximation (MLSA) (datorită răspândirii sale) şi Speech Transformation and Representation using Adaptive Interpolation of weiGHTed Spectrum (STRAIGHT) (Kawahara, 2006).

O alegere tipică pentru acest tip de modelare acustică sunt Modelele Markov Ascunse modificate să includă explicit durata ca parametru extern (fiind cunoscute sub numele de Hidden Semi-Markov Model HSMM), existând deja o serie de implementări directe de backend-uri de acest tip.

Pentru extinderea sistemului TTS RACAI, am ales utilizarea back-endului HMM Speech Synthesis System cunoscut şi sub numele HMM Tripple ‚S' (HTS). Pentru generarea unui model Statistic, HTS necesită un set de trăsături extrase pentru fiecare fonem aflat într-un context (propoziţie), fişierul acustic corespunzător (înregistrarea propoziţiei rostite) şi alinieri temporare între aceste secvenţe. Astfel, adaptarea sistemului TTS s-a efectuat prin realizarea unui adaptor de date responsabil pentru preluarea datelor generate de front-end şi conversia acestora într-un set de trăsături într-un format de intrare compatibil cu HTS. În plus, în momentul

antrenării, a fost necesară realizarea unui instrument de adăugare a informației de aliniere (extrase automat folosind HTK) între trăsăturile multinomiale şi fişierele acustice. Setul de trăsături ce trebuie extras nu este un corespondent direct al proceselor din frontend. Elementele de bază (transcriere fonetică, etichetare morfo-sintactică, despărțire în silabă etc.) trebuie îmbogățite cu informații ce aduc un aport modelării prozodice şi intonării corecte a propoziției date (lungimea propoziției, lungimea unui cuvânt, tipul propoziției – interogativă, declarativă, exclamativă etc.). Alegerea setului extins de trăsături influențează calitatea vocii sintetice (ajută la reducerea variabilității stărilor sistemului) şi sprijină modelarea prozodică. Extinderea setului de trăsături este de cele mai multe ori un proces euristic, care ține de dimensiunea corpusului de antrenare (pentru a evita supra-antrenarea sistemului) şi de stilul de vorbitorului (variabilitatea vocii sale din punct de vedere prozodic în diverse contexte ce țin de text şi semantică).

Deoarece îmbunătățirea predicției prozodice a sistemului TTS nu a intrat în scopul primei faze e proiectului, am ales un set de trăsături extinse care să constituie un suport preliminar pentru antrenarea sistemului şi realizarea prototipului experimental. Astfel, am ales să includem următoarea informație pentru fiecare fonem dedus din text: (a) Contextul fonetic (fonemele adiacente); (b) Silaba curentă; (c) Prezența accentului; (d) Numărul total de cuvinte din propoziție; (e) Numărul total de silabe din propoziție; (f) Numărul total de silabe accentuate din propoziție; (g) Distanța (în silabe) până la silaba anterioară accentuată; (h) Distanța (în silabe) până la silaba următoare accentuată; (i) Numărul de silabe accentuate înainte de silaba curentă; (j) Numărul de silabe accentuate după silaba curentă; (k) Tipul propoziției; (l) Poziția silabei curente față de începutul propoziției; (m) Poziția silabei curente față de finalul propoziției.

Realizarea unui modul experimental a atras după sine necesitatea existenței unui corpus de antrenare în vederea validării funcționale a prototipului. Una dintre resursele disponibile în acest sens este corpusul Romanian Speech Synthesis (RSS) (Stan et al., 2011). De asemenea, orice altă resursă audio ce este generată de un singur vorbitor, preferabil în condiții bune de înregistrare este binevenită în acest sens. Pentru o mai bună validare experimentală a prototipului am ales extinderea setului de date de intrare folosind înregistrări de tip audiobook.

Informația audio a fost extrasă dintr-o adaptare a cărții „În sfârşit nefumător" scrisă de Allen Carr. Corpusul este înregistrat de un vorbitor de gen masculin, în condiții de studio la o rată de eşantionare de 48 Khz. Cartea conține un număr mare de pasaje motivaționale şi persuasive are sunt atent construite de către autor pentru a convinge fumătorii să se lase de acest obicei. Astfel, vorbitorul a fost nevoit să recurgă la o vorbire cu elemente prozodice exagerate, recurgând deseori la retorică în vederea re-modelării emoționale a ascultătorului şi a introducerii de mesaje ascunse. Aceste lucruri ne-au determinat să considerăm această resursa foarte importantă în testarea şi validarea modelului experimental. Alinierea datelor s-a dovedit dificilă datorită numărului mare de abateri de la text a vorbitorului, generate fie de adaptări forțate (cuvântul „carte" a fost înlocuit sistematic cu „audiobook") fie de construcții artistice ale naratorului. În urma aplicării unui proceduri de filtrare a datelor (foarte sensibilă la erori de aliniere) am obținut un număr de 2.8K propoziții echivalente cu aproximativ 2 ore de vorbire continuă.

Modelul experimental obținut atât din corpusul RSS cât şi din datele extrase din audiobook este disponibil online[1]. Rata de vorbire precum şi timbrul vocal al vocii sintetice au fost schimbate pentru a nu semăna cu cele ale vorbitorului original.

4.2.3 Modelul experimental Voice Controller

Controllerul de voce este partea inteligentă a sistemului (vezi Fig. 1). În această etapă a fost construit un model experimental care are funcționalitatea descrisă mai jos. Controlerul de voce funcționează de inițiativă proprie la pornirea sistemului aplicând configurația implicită (default). În cursul funcționării normale, el are următoarele sarcini de realizat:

- să primească prin VoIP fluxul audio de la microfoanele instalate în casa inteligentă;
- să dirijeze fluxul audio către serverul ASR;
- să primească raspuns de la serverul ASR;

[1] http://rslp.racai.ro/index.php?page=tts

- să convertească răspunsul serverului ASR într-o comandă pentru sistemul casei automate (Building Operation System);
- să primească confirmare sau răspuns de la sistemul casei automate;
- să deducă răspunsul destinat utilizatorului din răspunsul sistemului casei automate;
- să trimită răspunsul text către serverul TTS;
- să primească fluxul audio de la sistemul TTS;
- să dirijeze fluxul audio primit de la sistemul TTS către difuzoare prin VoIP.

Sistemul permite şi primirea comenzilor prin telefon din exterior. În acest caz atribuțiile şi acțiunile controllerului de voce sunt aceleaşi cu cele descrise mai sus.

Prin intermediul controllerului de voce utilizatorul poate să işi definească toate comenzile ce se pot da sistemului de automatizare (atat setare-scriere cât şi status-citire).

În această etapă de model experimental am definit interfața cu utilizatorul ce permite configurarea sistemului şi interacțiunea dintre vocea persoanei şi sistemul de automatizare.

Interfata controllerului permite:

- Configurarea numărului şi parametrilor canalelor de comunicație cu serverul de VoIP.
- Configurarea comunicației cu ASR (parametrii).
- Configurarea comunicației cu TTS (parametrii).
- Trimiterea de comenzi către serverul de automatizare pe baza de cuvinte cheie şi parametri multipli.
- Formarea de fraze pe baza de parametrii returnați din serverul de automatizare şi sintetizate pe unul sau mai multe canale audio.

4.3 Proiectarea corpusului de vorbire (partea I) (Activitatea 1.3.)

Sistemele de sinteză a vorbirii bazate pe metoda concatenativă sau pe sinteză parametrică statistică cunosc o îmbunătățire a rezultatelor care depinde foarte mult de volum datelor de intrare. De asemenea, elementele prozodice învățate automat din corpus influențează modul de folosire a sistemului de sinteză a vorbirii. Astfel, un sistem de sinteză ce foloseşte un corpus din domeniu ştiri nu este uşor scalabil pentru a fi utilizat în citirea de nuvele sau basme. La fel de importantă este distribuția datelor de intrare: unitățile acustice de bază folosite în procesul de antrenare şi sinteză trebuie să apară cel puțin 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. Astfel, 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 se țină cont de următoarele cerințe: (a) corpusul trebuie să acopere mai multe domenii pentru a putea genera şi utiliza în funcție de caz un model corespunzător cu domeniul textului ce urmează a fi sintetizat; (b) unitățile 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.

Având în vedere aceste aspecte, a fost implementată o procedură automată menită ca, având un corpus mare de text, să aleagă propoziții distincte care să rezolve cerințele (b) şi (c). Sistemul de selecție a propozițiilor este unul deja consacrat (Matoušek et al., 2001), el fiind aplicat cu succes şi de către alte grupuri de cercetare. Procedura de selecție este următoarea:

1. În faza de pregătire a datelor, este ales un corpus de text care este pre-procesat folosind un front-end. Se urmăreşte astfel detectarea propozițiilor şi transcrierea fonetică a cuvintelor;
2. Se creează o bază de propoziții unice şi se alege un număr minim dorit de ocurențe pentru fiecare tri-fonem;
3. La fiecare pas, se alege o propoziție care conține cât mai multe tri-foneme ce au un număr mic de apariții în propozițiile deja selectate;
4. Condiția de oprire este dată de atingerea numărului minim dorit de ocurențe pentru fiecare tri-fonem sau de epuizarea datelor de intrare.

Având în vedere modul în care acest corpus va fi folosit în viitorul proiectului ANVSIB dar şi ulterior, domeniile de interes principal vor fi ştiri şi romane moderne. În ambele cazuri, se vor impune restricții referitoare la lungimea unei propoziții preluate şi înregistrate, pentru a fi în conformitate cu reglementările

privind drepturile de autor. În afară de domeniile mari de interes, corpusul va conține o componentă de bază acustică neutră care va fi obținută din Wikipedia.

4.4 Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteză şi recunoaştere a vorbirii (Activitatea 1.4.)

Proxy-ul de voce are rolul de a fi interfața între utilizatori şi cele 2 sisteme de voce (ASR sau TTS) aşa cum este arătat în Fig. 1. Acest modul a fost introdus pentru a respecta principiul de interoperabilitate şi independență între sisteme. S-a folosit o tehnologie solidă de VoIP pentru vehicularea fluxului audio. Această tehnologie permite unificarea modului de comunicație între difuzoare şi controllerul de voce şi între telefoane şi controllerul de voce.

Utilizatorul, pentru a avea mobilitate, se poate folosi de terminale clasice compatibile cu protocolul de comunicație SIP dar şi de telefoane mobile/tablete sau orice alt gateway SIP. Prin intermediul protocolului MRCP (Media Resource Control Protocol) şi SIP acesta este capabil să ofere mobilitate utilizatorului pentru a putea interpreta comenzile şi a sintetiza răspunsuri la intrebarile acestuia.

Initial am luat în calcul conectarea fluxurilor audio direct la sistemele TTS şi ASR, pentru a evita dublarea traficului audio pe infrastructura de rețea. Având în vedere că anumite comenzi trebuie sintetizate pe toate canalele audio dintr-o clădire (de exemplu: "Atenție! Se armează sistemul de efracție") am decis ca toate fluxurile audio să fie gestionate de voice controller pentru a le putea ruta. Am ajuns la concluzia ca interfața controllerului trebuie să ofere posibilitatea definirii numarului de dispozitive/gateway-uri pentru transportul audio, dar şi posibilitatea de a te conecta din exteriorul casei.

Voice controllerul se conectează la serverul ASTERISK şi are posibilitatea transportării fluxurilor audio atât din interiorul casei cât şi din exteriorul ei dacă serverul de ASTERISK este interfațat cu o centrală clasică PBX sau printr-o conexiune IP şi un client VOIP pe un terminal mobil (Android/IOS).

4.5 Achiziționarea unei baze de date de vorbire in conditii reale (partea I) (Activitatea 1.5.)

Bazele de date de înregistrări audio sunt esențiale în dezvoltarea sistemelor de recunoaştere automată a vorbirii. Acestea sunt utilizate atât în procesul de antrenare pentru a crea modele acustice ce au rolul de a caracteriza modul de pronunție al sunetelor de bază ale unei limbi, cât şi în procesul de evaluare pentru stabilirea acurateții sistemului de recunoaştere automată a vorbirii.

În cadrul activității 1.5 a fost creată o bază de date de înregistrări audio pe un grup de propoziții pronunțate în condiții reale (fiecare vorbitor utilizând propriul dispozitiv de înregistrare). Fiecare dintre vorbitori a înregistrat întregul set de fraze folosind aplicația dezvoltată şi pusă la dispoziție de Universitatea Politehnica Bucureşti – UPB. Utilitarul folosit pentru înregistrări se găseşte online, iar instrucțiunile pentru rularea înregistrărilor se găsesc aici.

Paşii urmați pentru efectuarea înregistrărilor au presupus:

- accesul online la pagina utilitarului
- rularea testului online pentru calibrarea microfoanelor
- verificarea calității înregistrărilor

În Fig. 2 este prezentată o imagine din cadrul aplicației online de înregistrări.

Figura 2. Aplicația online de înregistrări

Baza de date de comenzi a fost creată urmărind lista de sisteme configurabile întocmită de partenerul P1 (iWave Solutions). Frazele au fost selectate având în vedere următoarele aspecte:

- frazele ar trebui să fie corecte din punct de vedere gramatical;
- frazele ar trebui să conțină toate semnele de punctuație (pe care vorbitorul le va utiliza pentru a citi cu intonație);
- numerele din fraze ar trebui să fie scrise cu cifre pentru a lăsa vorbitorului libertatea de a-şi alege singur modul de pronunție, fără a-l limita în vreun fel;
- frazele ar trebui să conțină numai comenzi pe baza listei de sisteme configurabile;
- frazele ar trebui să fie scurte pentru a putea fi pronunțate fără ezitări sau respirații;
- setul de fraze ar trebui să conțină şi fraze interogative, exclamative, dialog;
- frazele ar trebui să conțină diverse moduri de exprimare pentru o aceeaşi comandă (ex:”Casandra, pornește stropitorile”, “Casandra, dă drumul la stropitori”) cu scopul de a acoperi o arie cât mai mare a vorbirii colocviale
- frazele ar trebui să fie concise pentru a nu adăuga informație inutilă în partea de recunoaştere

Baza de date este compusă din înregistrările a 5 vorbitori (3 voci masculine şi 2 voci feminine), fiecare dintre vorbitori înregistrând întregul set de fraze. Setul de comenzi constă în 204 comenzi, fiecare dintre ele fiind precedată de cuvântul cheie “Casandra”, totalizând un număr de 963 de cuvinte. Această bază de date a fost creată pentru a testa funcționalitatea sistemului ASR, fapt ce motivează dimensiunea sa redusă. În dezvoltarea viitoare a sistemului este prevăzută şi achiziționarea unei baze de date mai amplă atât din punctul de vedere al numărului de vorbitori cât şi din punctul de vedere al numărului de comenzi.

Toate înregistrările audio şi transcrierile aferente au fost organizate într-un director cu următoarea structură:

/wav	Subdirector în care sunt grupate toate fişierele audio (în format .wav, 16kHz, 16 biți/eşantion)
/wav/<speaker-id>	Subdirectoare specifice fiecărui vorbitor în care sunt grupate toate fişierele audio ale respectivului vorbitor. Înregistrările

	audio au fost realizate de 5 de vorbitori diferiți cu id-urile: 139, 264, 300, 319 şi 356.
/wav/<speaker-id>/<recording-id>.wav	Fişiere audio înregistrate de vorbitorul cu id-ul <speaker-id>. Denumirile fişierelor audio (<recording-id>.wav) respectă formatul: AAA_BB_CCCC.wav, unde AAA este id-ul vorbitorului, BB este id-ul grupului de fraze, iar CCCC este id-ul înregistrării.
/transcription	Subdirector în care sunt grupate toate transcrierile fişierelor audio
/transcription/<phrase-group>.transcription	Fişiere text ce conțin transcrierile tuturor frazelor din grupul de fraze cu id-ul <phrase-group>. În cazul acestei baze de date <phrase-group> = 41.

4.6 Diseminarea rezultatelor proiectului (partea I) (Activitatea 1.6.)

În cadrul acestei activități a fost creată pagina web a proiectului care conține informații cheie cum ar fi:
- finanțatorul proiectului;
- descrierea partenerilor şi rolul lor în proiect;
- rezumatul proiectului;
- durata proiectului.

Proiectul a fost făcut cunoscut în forumurile în care laboratorul SpeD are acces.
Adresa paginii web este: **http://speed.pub.ro/anvsib**

4.7 Sumar al realizărilor

Nr.	Denumirea rezultatului	Denumire activitate	Procent de realizare
1.	*Set de cerințe arhitecturale*	*1.1 Cerințe arhitecturale*	*100%*
2.	*Model experimental ASR* *Model experimental TTS* *Model experimental Voice Controller*	*1.2 Modele experimentale pentru componentele sistemului*	*100%*
3.	*Colecție de propoziții*	*1.3 Proiectarea corpusului de vorbire (partea I)*	*100%*
4.	*Proxy de voce*	*1.4 Crearea proxy-ului de voce pentru integrarea VoIP a sistemelor de sinteza si recunoastere a vorbirii (partea I)*	*100%*
5.	*Baze de date de înregistrări audio*	*1.5 Achiziționarea unei baze de date de vorbire in condiții reale (partea I)*	*100%*
6.	*Pagină web a proiectului*	*1.6 Diseminarea rezultatelor proiectului (partea I)*	*100%*

5 Documente în extenso

Nr.	Denumire anexă	Activitate asociata	Tip
1.	FSG_Casandra_v3.gram	1.2. Modele experimentale pentru componentele sistemului	Model lingvistic de tip gramatica cu stari
2.	CommandsList_v4.txt	1.5. Achiziționarea unei baze de date de vorbire in condiții reale (partea I)	Definiție listă comenzi

6 Rezultate

Nr. Crt.	Denumire	Tip	An obtinere
1	**Modul experimental pentru ASR**	**Produs informatic**	**2014**
2	**Modul experimental pentru TTS**	**Produs informatic**	**2014**
3.	**Modul experimental pentru Voice Controller**	**Tehnologie**	**2014**
4.	**Pagină web a proiectului**	**Altele**	**2014**

Referințe bibliografice

Bartlett, S., Kondrak, G., & Cherry, C. (2008). Automatic syllabification with structured SVMs for letter-to-phoneme conversion. Proceedings of ACL-08: HLT, 568-576.

Boroş, T., Ştefănescu, D., Ion, R.(2013a). Handling Two Difficult Challenges for Text-to-Speech Synthesis Systems: Out-of-Vocabulary Words and Prosody -- A Case Study in Romanian, book chapter, Where Humans Meet Machines, edited by Amy Neustein, Springer, 2013.

Boroş, T., Ion, R., Tufiş, D.(2013). Large tagset labeling using Feed Forward Neural Networks. Case study on Romanian Language, accepted for publication in ACL, Sofia, Bulgaria, 2013.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustical science and technology, 27(6), 349-353.

Matoušek, J., Psutka, J., & Krůta, J. (2001). Design of speech corpus for text-to-speech synthesis.

Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Communication, 53(3), 442-450.