REALIZACIJA TELEFONSKOG GOVORNOG PORTALA NA BAZI ASR I TTS

Robert Ronto, Darko Pekar, *AlfaNum – govorne tehnologije, Novi Sad*
Nikola Đurić, *Fakultet tehničkih nauka, Novi Sad*

Sadržaj – *Interaktivni govorni portal za slepe i slabovide „Kontakt" realizovan je kao modularni objedinjeni telefonski i web portal sa mogućnošću proširivanja koji nudi sadržaje osobama oštećenog vida preko javne telefonske mreže i Interneta putem govora. Tema ovog rada jesu rešenja primenjena u implementaciji konverzije teksta u govor (TTS) i automatskog prepoznavanja govora (ASR) pri realizaciji telefonskog dela govornog portala.*

1. UVOD

Govorni portal "Kontakt" razvijen je za PC platformu sa Dialogic CTI (Computer Telephony Integration) karticom Springware tehnologije [1] koja vrši obradu govornog signala i komunikaciju PC platforme sa telefonskom linijom, čime je realizovan telefonski deo portala, i Internet Information Serverom uz PHP i MySQL [2] koji obavlja hostovanje web portala na Internetu. Portal se trenutno nalazi u fazi testiranja, na serveru u laboratoriji za govorne tehnologije Fakulteta tehničkih nauka u Novom Sadu. I web i telefonski portal oslanjaju se na jedinstvenu bazu podataka kao izvor korisnog sadržaja. Baza trenutno sadrži mnoštvo informacija, kako iz oblasti informisanja (vesti, pravna akta od interesa), tako i iz oblasti zabave (repertoari pozorišta, koncerti, sport), kao i mogućnost teoretski neograničenog proširenja i koncipiranja po želji administratora.

Informacijama u bazi moguće je pristupiti preko interneta (na adresi http://www.alfanum.ftn.ns.ac.yu/kontakt) kao običnoj web strani koristeći web browser, čiji se sadržaj nekim od čitača ekrana (npr. Jaws) pretvara u govor, omogućujući na taj način efikasno korišćenje internet portala slepim i slabovidim korisnicima.

Telefonski portal realizovan je kao govorni automat koji govornim menijima, identičnim menijima sa linkovima na internet portalu, omogućuje navigaciju i preslušavanje sadržaja zadavanjem govornih komandi. Da bi korisnik portala pristupio sadržaju potrebno je da govornom navigacijom kroz niz intuitivnih menija izabere željenu oblast koja sadrži informacije od interesa.

Da bi se korisnik lakše kretao po granama menija, svaka stavka menija podeljena je na dva dela: prvi deo, nazvan TEME, sadrži nazive podmenija trenutne stavke, dok drugi deo, nazvan NASLOVI sadrži naslove raspoloživih sadržaja koji se mogu iščitati, pri čemu se pod „iščitavanjem" podrazumeva sinteza teksta iz baze podataka u govor. Svaka stavka, tj. grana menija, analogno internet potalu, naziva se STRANA.

Korisnik komandom TEME bira iščitavanje naziva podmenija koji su mu na raspolaganju, a izgovaranjem naziva podmenija ostvaruje pomak u strukturi menija, tako da mu se iščitava naslov trenutne (odabrane) stavke/strane i očekuje nova komanda.

Izgovaranjem komande NASLOVI korisniku se iščitavaju naslovi sadržaja koji su mu na raspolaganju. Validne komande su PRVI (iščitavanje počev od prvog naslova), POSLEDNJI (iščitavanje poslednjeg naslova), SLEDEĆI / DALJE i PRETHODNI (iščitavanje sledećeg i prethodno iščitanog naslova, respektivno).

Sam sadržaj iščitava se komandom ČITAJ izrečenom neposredno posle naslova. Čitanje se prekida komandom STOP ili iščitavanjem celokupnog sadržaja, posle čega se nastavlja iščitavanje naslova počevši od naslova koji sledi posle naslova iščitanog sadržaja. Pri iščitavanju sadržaja, validne komande su POČETAK (sadržaj se čita od početka), NAPRED (čita se sledeća rečenica), i NAZAD (čitanje prethodne rečenice).

Na nivou strane, komandom POČETNA, korisnik se može vratiti na uvodnu stranu, a komandom NAZAD kreće se na prethodnu savku menija.

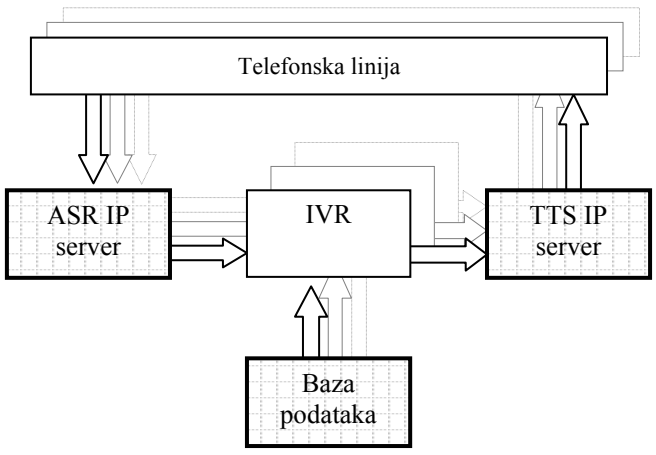
Treba napomenuti da za strane koje sadrže isključivo oblasti ili isključivo sadržaje ne zahtevaju komandu TEME/NASLOVI, već se iščitavanje podmenija (tema) ili naslova sadržaja vrši automatski.

2. TELEFONSKI GOVORNI PORTAL "KONTAKT"

Celokupan govorni portal oslanja se na bazu podataka kao izvor sadržaja. Da bi se korisniku saopštio traženi sadržaj, potrebno je postaviti upit bazi podataka u zavisnosti od sadržaja zahtevane oblasti. Traženi sadržaj jeste rezultat upita iz baze podataka.

Kako bi sistem bio dovoljno efikasan, potrebno je da omogući istovremen pristup različitim sadržajima dovoljnom broju korisnika.

Ulaz u sistem telefonskog govornog portala jeste ljudski govor. Da bi se iz ulaza u sistem izdvojila korisna informacija potrebno je ljudski govor pretvoriti u prepoznatljiv nosilac podataka za računarski sistem, u konkretnom slučaju u string ili niz stringova karaktera koji su pogodni za formiranje upita u bazu podataka. Isto tako,



Sl.1. *Komunikacija u sistemu telefonskog govornog portala*

rezultat upita jeste string karaktera, koji je, da bi predstavljao

upotrebljiv izlaz iz sistema na telefonsku liniju, potrebno pretvoriti u ljudski govor. Sa druge strane, da bi sistem zadovoljio zahteve u pogledu efikasnosti, potrebno je da funkcioniše na više telefonskih linija istovremeno i nezavisno od linije do linije.

Uzevši u obzir resurse računarskog sistema i navedene zahteve, rešenje koje je primenjeno sastoji se iz IVR (Interactive Voice Response) aplikacije koja opslužuje po jednu telefonsku liniju u sprezi sa ASR i TTS serverima koji preko IP protokola komuniciraju sa potrebnim brojem IVR aplikacija, i jedinstvene baze podataka. Prednost ovakvog pristupa ogleda se u činjenici da ASR i TTS serveri mogu biti locirani i na udaljenim računarima koji mogu biti posvećeni isključivo ASR i/ili TTS konverzijama, i da takvi serveri mogu komunicirati sa većim brojem različitih IVR aplikacija. Mana bi bila usporen odziv (nedovoljna brzina konverzije) u slučaju preopterećenosti servera zbog prevelike količine podataka za konverziju prouzrokovane velikim brojem aplikacija i/ili predugačkim ulaznim sekvencama.

U realizovanom sistemu korišćeni su AlfaNum ASR IP server [3] i AlfaNum TTS IP server [4], razvijeni u okviru projekta AlfaNum na Fakultetu tehničkih nauka u Novom Sadu.

Sadržaj portala pohranjen je u MySQL bazu podataka što omogućuje standardizovano generisanje web strana PHP skriptovima, kao i pouzdanu vezu sa IVR aplikacijom koja je realizovana preko C API-a MySQL-a.

Baza podataka osvežava se puller aplikacijom koja je koncipirana tako da nove sadržaje periodično prikuplja sa Interneta. Sadržaji se administriraju sa web strane portala, preko običnog browsera, prijavom na portal sa korisničkim imenom administratora i odgovarajućom lozinkom.

Komunikacija sistema sa telefonskom linijom vrši se preko Dialogic CTI kartice, a kontrola komunikacije korišćenjem Dialogic dx i Global Call API-a. Broj linija koje sistem opslužuje zavisi od vrste i broja Dialogic kartica integrisanih u PC platformu sistema.

3. ASR SERVER

Kao što je napomenuto, ASR server vrši pretvaranje ljudskog govora u niz tekstualno notiranih reči, tj. stringova, koji su prihvatljivi nosilac podataka za računarski sistem. Sam proces pretvaranja naziva se prepoznavanje govora.

Da bi prepoznavanje uspešno funkcionisalo, potrebno je definisati skup reči koje ASR server može očekivati na svom ulazu u određenom trenutku. Pravila po kojima server vrši prepoznavanje u različitim vremenskim trenucima definišu niz prepoznavača u okviru ASR servera. Na primer, početna strana portala sadrži oblasti i sadržaje, tako da je potrebno odabrati jednu od te dve podkategorije, što korisnik sistema čini komandom TEME ili NASLOVI. Znači, u trenutku kada se govorni automat nalazi u stanju početnog menija, prepoznavač očekuje reči TEME ili NASLOVI. Pravila po kojima se vrši takvo prepoznavanje definišu prepoznavač, a svaki prepoznavač se opisuje gramatikom. Pri svakom prepoznavanju ulaz u prepoznavač je govorni signal koji se prepoznaje i gramatika koja definiše pravila prepoznavanja, tj. skup očekivanih reči i način njihovog pojavljivanja. Po izvršenom prepoznavanju izlaz ASR servera predstavlja niz prepoznatih reči predstavljen kao niz stringova i niz cifara koji predstavlja stepen pouzdanosti kojom je prepoznata svaka reč prvog niza.

Niz cifara pouzdanosti je veoma važna izlazna veličina prepoznavača. Kako su gramatikama strogo definisana pravila prepoznavanja, prepoznavač nije u stanju da ispravno procesira reči van skupa definisanog u gramatici koja se koristi u datom trenutku, tako da u slučaju govornog signala na ulazu koji nije obuhvaćen gramatikom, na izlazu ipak biti neka od reči definisanih u gramatici, ali sa veoma malim stepenom pouzdanosti, tako da se pravilnim podešavanjem praga pouzdanosti neminovna greška može izbeći: ulaz se deklariše kao nevalidan, te se zahteva ponovno izgovaranje reči za prepoznavanje.

Gramatike su reprezentovane tekstualnim fajlovima koji se učitavaju pri startovanju ASR servera. Svaki ulaz u prepoznavač može biti definisan različitim gramatikama, što u slučaju telefonskog portala i jeste slučaj. Primer jedne gramatike, konkretno one koja opisuje ulaz prepoznavača u početnom stanju menija bi izgledao ovako:

```
cmd = TEME | NASLOVI;
gr = <gar>;
main = [$gr] [$cmd] [$gr];
```

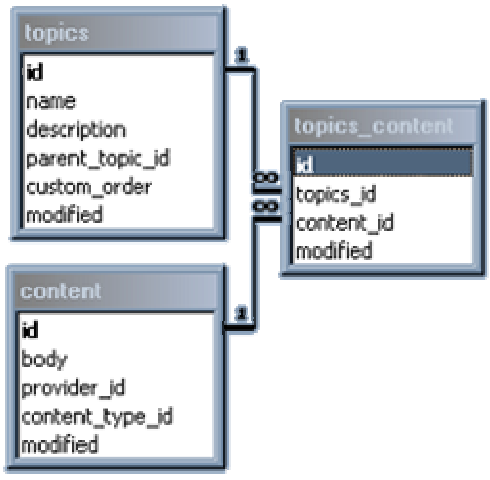
gde `<gar>` označava proizvoljnu smetnju koja se pri prepoznavanju zanemaruje.

Da bi korisnik vršio uspešnu navigaciju kroz govorni meni, potrebno je za svaki (pod)nivo menija definisati gramatiku za prepoznavač. Struktura menija zavisi isključivo od definisanih oblasti (tema) u bazi podataka, a jedan od projektnih zadataka jeste da sadržaj baze bude proizvoljno proširiv i pogodan za modifikovanje, tako da ne ograničava proizvoljno koncipiranje portala. Pri tako definisanim zahtevima, javlja se problem koji se ogleda u nemogućnosti predefinisanja gramatika prepoznavača.

Jedino prihvatljivo rešenje jeste automatsko generisanje gramatika iz baze podataka. Na ovom stepenu razvoja sistema, gramatike se generišu pri startovanju IVR aplikacije, posle čega se startuje ASR server, tako da je prepoznavač pri svakom startu inicijalizovan sa aktuelnim gramatikama. Jedina mana ovakvog pristupa ogleda se u činjenici da je neophodno restartovati ASR server ako se osvežavanje baze podataka izvrši za vreme rada aplikacije.

3.1. GENERISANJE GRAMATIKA

Tabele baze podataka koje su relevantne za formiranje



Sl.2. *Tabele za generisanje menija i gramatika*

menija i generisanje gramatika prikazane su na slici 2.

Tabela topics sadrži sve oblasti, tj. teme po kojima su sadržaji grupisani, a u tabeli content smešteni su sadržaji i relevantni podaci o njima (izvor/autor, tip, datum). Tabela topics_content povezuje teme sa sadržajima, formirajući grupe sadržaja po temama.

Svaka tema (tabela topics) pripada određenoj temi koja se nalazi neposredno iznad nje u hijerarhiji menija. Ova pripadnost definiše se poljem parent_topic_id – identifikatorom teme kojoj aktuelna tema pripada. Teme koje ne pripadaju drugim temama (nalaze se na vrhu stabla menija) imaju parent_topic_id jednak nuli.

Formiranje govornog menija vrši se na sledeći način: označimo stavku menija, tj stranu koja se u nekom trenutku iščitava jednoznačnim identifikatorom sa nekom vrednošću *id*. Podmeniji, tj. stavke menija koje pripadaju trenutnoj strani su sve teme iz tabele topics čija je vrednost polja parent_topic_id jednaka vrednosti *id*. Sadržaji koji su uključeni u trenutnu stavku menija definisani su u tabeli topics_content identifikatorom topics_id koji je takođe jednak vrednosti *id*. Znači, da bi se formirale grane menija i sadržaji koji su na raspolaganju korisniku u datom trenutku, potrebno je izvršiti upit u bazu podataka za svim temama (tabela topics) i sadržajima (tabela content u sprezi sa tabelom topics_content) čiji identifikatori pripadnosti (parent_topic_id za teme i topics_id u tabeli topics_content za sadržaje) imaju vrednost jednaku identifikatoru trenutne strane *id*.

Pri generisanju gramatika situacija je pojednostavljena utoliko što iščitavanje sadržaja ne zahteva formiranje posebne gramatike, pošto su komande ČITAJ, SLEDEĆI/DALJE, PRETHODNI, PRVI i POSLEDNJI zajedničke za sve strane portala, što proizilazi iz primenjenog rešenja koje je opisano u uvodu.

Potrebno je, međutim, generisati gramatike za raspoložive teme na strani, kako bi se obezbedilo prepoznavanje za navigaciju, tj. grananje menija. Generisanje gramatika vrši se po rezonu identifikatora pripadnosti, analogno formiranju govornog menija. Za svaku stranu portala generiše se po jedna gramatika koja iz baze definiše pristupačne teme, uz konstante prepoznavanja, koje predstavljaju reči koje su prisutne u svim gramatikama, a obezbeđuju neke spacifične akcije proistekle iz dizajna portala. Specifične akcije predstavljaju npr. povratak na prethodnu stranu (nazad), povratak na početnu stranu (početna), traženje uputsva (pomoć) ili podešavanje parametara sinteze govora (opcije), o čemu će biti reči u sledećem poglavlju.

Da bi se generisanje gramatike uprostilo, svaka generisana gramatika kao zajedničku polaznu tačku ima predefinisanu inicijalnu gramatiku. Inicijalna gramatika sadrži elemente koji se pojavljuju u svakoj od njih, kao što su kontante prepoznavanja i struktura koja se upotrebljava.

Deo gramatičkog fajla koji je promenjiv ograničen je stringovima `/*genericstart*/` i `/*genericend*/`, kako bi se pri generisanju gramatike iz inicijalne ili eventualnom ponovnom generisanju iste gramatike, koja bi bila neophodna u slučaju promena u bazi podataka, menjao samo tzv. dinamički deo gramatike, dok bi konstante i zajednička struktura ostale nepromenjene. Napomenimo da karakteri `/*` i `*/` sa gledišta prepoznavača uokviruju komentar, tako da se uokvireni sadržaj ne procesira, dok za parser IVR aplikacije stringovi između spomenutih karaktera definišu početak i kraj dela fajla koji je potrebno (re)generisati.

Primer inicijalne gramatike koji sadrži konstante prepoznavanja NAZAD, POMOĆ i OPCIJE:

```
navig = NAZAD | POMOCC | OPCIJE;
/*genericstart*/
/*genericend*/
komanda = $navig | $generic;
gr = <gar>;
main = [$gr] [$komanda] [$gr];
```

Primer jedne jednostavne gramatike generisane iz gornje inicijalne gramatike:

```
navig = NAZAD | POMOCC | OPCIJE;
/*genericstart*/
koncerti = KONCERI;
pozorishta = POZORISHTA;
generic = $koncerti | $pozorishta;
/*genericend*/
komanda = $navig | $generic;
gr = <gar>;
main = [$gr] [$komanda] [$gr];
```

Svaki generisani fajl gramatike naziva se po identifikatoru strane na kojoj se upotrebljava. Na taj način je uprošćena i generalizovana upotreba konkretnog prepoznavača iz koda IVR aplikacije.

Da bi se generisane gramatike uspešno koristile od strane ASR servera, potrebno je lokaciju svakog novog fajla koji predstavlja gramatiku na neki način saopštiti serveru. To se čini preko inicijalizacionog fajla ASR servera koji sadrži sva potrebna podešavanja koja definišu rad samog servera, kao što su parametri obrade govornog signala, koeficijenti vezani za samo prepoznavanje, IP port na kojem server vrši komunikaciju i podaci IP servera. U podatke IP servera spada vektor prepoznavača (recognizers) koji za svaki prepoznavač definiše ime, putanju fajlova gramatike, postprocesora, rečnika izgovora pojedinih reči i transkriptora (opšteg skupa pravila fonetske transkripcije). Dakle, da bi se ASR server uspešno inicijalizovao novonastalim gramatikama, potrebno je za svaku od njih formirati prepoznavač sa svim potrebnim parametrima i uvrstiti ga u vektor prepoznavača.

Znači, kao što je ranije spomenuto, pod prepoznavačem se podrazumeva skup pravila po kojima server vrši prepoznavanje u konkretnom trenutku vremena. Sa aspekta IVR aplikacije, prepoznavanje u određenom vremenskom trenutku definiše se imenom konkretnog prepoznavača iz vektora prepoznavača ASR servera po čijim pravilima se želi vršiti prepoznavanje.

4. TTS SERVER

Sinteza teksta u govor neophodna je kako bi se korisna informacija iz baze podataka efikasno prenela do svog odredišta (korisnika sistema) preko telefonske linije. Uz podatke iz baze u govor je neophodno pretvoriti i sve povratne informacije iz sistema, kao što su meni za navigaciju, trenutna lokacija korisnika u stablu menija (trenutna strana), raspoložive oblasti i sadržaji na trenutnoj strani, parametri sinteze govora po potrebi, itd.

Sinteza teksta vrši se upotrebom `synthesize` metode TTS servera. Ulazna veličina metode je tekst, tj. string koji se želi pretvoriti u govor, dok je izlazna veličina .wav fajl koji sadrži ulazni tekst u govornoj formi. Komunikacija sa TTS serverom (slanje ulaznog stringa i prijem .wav fajla) vrši se posredstvom IP protokola.

Trajanje procesa sinteze zavisi od veličine ulaznog teksta, a vreme koje je potrebno za sintezu za korisnika sistema predstavlja vreme neaktivnosti sistema, tj. vreme provedeno u čekanju odziva govornog portala. Za ulazne stringove reda veličine rečenice, kao što su meniji, naslovi sadržaja, broj oblasti ili sadržaja na strani, saopštavanje parametara sintetizatora, vreme sinteze je prihvatljivo i ne predstavlja značajno kašnjenje sistema, tako da metoda `synthesize` zadovoljava zahteve u pogledu brzine konverzije.

Problem se, međutim, javlja pri sintezi konkretnog sadržaja iz baze podataka sistema. Pošto je portal koncipiran tako da ne ograničava ni sadržaj, ni vrste sadržaja pohranjenih u bazu, dužina teksta koji predstavlja konkretan sadržaj je praktično neograničena i poseduje red veličine daleko iznad jedne rečenice. Kao ilustraciju, navedimo primer da se trenutno u bazi nalaze sadržaji kao što su novinski članci, mesečni repertoari pozorišta, pa i pravna akta. Sinteza ovakvih sadržaja metodom `synthesize` trajala bi neprihvatljivo dugo i učinila sistem neupotrebljivo sporim.

Da bi se ubrzala sinteza velike količine teksta implementirana je sinteza toka (stream) teksta koja se bazira na kontrolisanoj kontinualnoj sintezi blokova ulaznog niza stringova. Suština takve sinteze jeste rasparčavanje ulaznog teksta na manje delove (blokove) koji obezbeđuju zadovoljavajuću brzinu sinteze i puštanje sintetizovanog govora korisniku za vreme trajanja sinteze sledećeg bloka ulaznog teksta. Kontrolisana sinteza toka teksta postiže se umetanjem kratkog vremena čekanja između sinteze blokova teksta koje obezbeđuje prepoznavanje komandi korisnika i omogućuje upravljanje sintezom u smislu sintetisanja prethodnog/sledećeg bloka, sintetisanja od početka teksta ili prekid sinteze.

U sistem je implementirana sinteza toka teksta razvijena za potrebe aplikacije za čitanje elektronske pošte anMailReader kompanije Alfanum, a više detalja vezanih za sintezu kontinualnog toka teksta može se naći u [5].

4.1. PARAMETRI TTS KONVERZIJE

Alfanum TTS IP server pruža mogućnost sinteze teksta u govor čiji su parametri, kao što su visina glasa, brzina izgovaranja i pol govornika, konfigurabilni. U sadašnjem stepenu razvoja TTS server poseduje dva govornika, muškog (koji je nazvan Steva) i ženskog (Marija).

Da bi sistem omogućio ugodnije korišćenje i podešavanje parametara sintetizovanog govora po želji korisnika, uvedena je komanda OPCIJE, kojom se inicira govorni meni za podešavanje parametara sinteze.

Podmeni OPCIJE sadrži tri stavke od kojih svaka obezbeđuje promenu po jednog od navedenih parametara TTS servera, a to su: GOVORNIK, VISINA i BRZINA.

Napomenimo da se promena parametara sinteze može izvršiti korak po korak navigacijom kroz podmenije menija OPCIJE, ili spajanjem komandi izbora u jednu frazu, na primer GOVORNIK MARIJA, što omogućuje bržu konfiguraciju iskusnijim korisnicima.

Dizajnom baze predviđeno je i pamćenje profila korisnika, tako da se prijavom korisnika na servis parametri sinteze automatski podešavaju na vrednosti koje je korisnik podesio tokom prethodne sesije.

5. ZAKLJUČAK

Razvoj telefonskog govornog portala "Kontakt" omogućio je bolji uvid u problematiku realizacije CTI aplikacija koje integrišu postojeća ASR i TTS rešenja. Razvijeni test sistem ukazuje na moguća ograničenja ASR/TTS modula koji su implementirani kao IP serveri, a test faza eksploatacije sistema pokazuje optimalan odnos opterećenosti sistema i potrebnog broja servera u zavisnosti od brzine upotrebljene platforme. Javlja se potreba za daljim razvojem i optimizacijom sinteze kontinualnog toka velike količine tektualnih sadržaja radi komfornijeg korišćenja ovakvih sistema, kao i buduća integracija takve sinteze u postojeći TTS IP server.

Implementacijom ASR u sprezi sa dizajnom baze podataka kakav je primenjen u sistemu postignuta je primena prepoznavanja govora koja nije ograničena sadržajem, tj. ne mora biti predefinisana za vreme dizajna i /ili razvoja sistema, što u perspektivi širi spektar upotrebe prepoznavanja govora u telefonskim ili sličnim aplikacijama.

ZAHVALNICA

Ovaj rad je podržan od Ministarstva nauke i zaštite životne sredine u okviru inovacionog projekta (PTR 2079A) pod nazivom "Govorni portal za slepe i slabovide na srpskom govornom području – Kontakt".

LITERATURA

[1] http://resource.intel.com/telecom/support/releases/winnt/ /SR511/docs/htmlfiles/index.html

[2] J. Greenspan, B. Bulger, *MySQL/PHP database applications*, IDG Books, 2001

[3] D. Pekar, R. Obradović, *Programski paket AlfaNumCASR - sistem za prepoznavanje kontinualnog govora*, DOGS 2002

[4] M. Sečujski, R. Obradović, *Alfanum sistem za sintezu govora na osnovu teksta na srpskom jeziku*, DOGS 2002

[5] Lj. Jovanov, D. Pekar, *AnMailReader - CTI aplikacija za čitanje e-mail poruka*, DOGS 2004

Abstract – „Kontakt" is an interactive voice portal for visually impaired people and is developed as a modular, upgradeable, unified telephone and web portal. It offers various contents by means of speech for the visually impaired accessible over public telephone network and the Internet. The purpose of this paper is to discuss the text to speech synthesis (TTS) and automatic speech recognition (ASR) methods used in the development of the telephone portion of the portal.

DEVELOPING A TELEPHONE VOICE PORTAL WITH ASR AND TTS CAPABILITY

Robert Ronto, Darko Pekar, Nikola Đurić