Univerza v Ljubljani
Naravoslovnotehniška fakulteta
Oddelek za tekstilstvo
Grafična tehnologija

Seminarska naloga pri predmetu Jezikovne tehnologije

SpeechDat II

Mihaela Rožej
in
Nadja Srebot

Ljubljana, 2006

Kazalo

1 Uvod

Večina obstoječih avtomatsko govorno krmiljenih telefonskih storitev uporablja razpoznavanje osamljenih besed. Da bi povečali prijaznost takih sistemov in jim zagotovili konkurenčnost tudi v prihodnje, je potrebno razviti sisteme, ki uporabljajo razpoznavanje tekočega ali celo spontanega govora. Za razvoj takih sistemov so potrebne obsežne govorne baze. Te morajo vsebovati tako besede, specifične za aplikacijo, če je ta v naprej znana, kakor tudi fonetično bogate besede in stavke, namenjene za razvoj novih aplikacij. Z vključitvijo fonetično bogatega teksta se izognemo potrebi po zbiranju baze za vsako novo aplikacijo, ki jo razvijamo.

Slovenska SpeechDat baza izgovorjav za stacionarno telefonsko omrežje je bila izdelana z namenom zapolniti praznino, ki je zevala na področju govornih baz za slovenščino. Posneta je bila v okviru projekta SpeechDat II (LE2-4001), ki ga je sponzorirala EU in vsebuje posnetke 1000 govorcev. Baza je bila zasnovana in posneta na Univerzi v Mariboru v Laboratoriju za digitalno procesiranje signalov Fakultete za elektrotehniko, računalništvo in informatiko. Projekt izdelave baze je finančno podprl Siemens A.G., ki tudi koordinira SpeechDat projekt.

2 Zasnova podatkovne baze in zbiranje podatkov

2.1 Snemalno okolje in anotacija

Baza je bila posneta preko stacionarnega telefonskega omrežja. Snemalni strežnik je bil stacioniran v Laboratoriju za digitalno procesiranje signalov na Univerzi v Mariboru. Na telefonsko omrežje je bil priključen preko ISDN linije. Strežnik je sestavljal PC računalnik z ISDN telefonsko kartico. Posnetke so tedensko prenašali na delovne postaje, kjer so bili anotirani in dokončno obdelani.

2.2 Pridobivanje govorcev

Govorci v bazi so bili večinoma zaposleni na Pošti Slovenije. Izbrani so bili glede na postavljene zahteve za uravnoteženost porazdelitve govorcev po spolu, starosti in narečju. Ker ima Pošta urade po vsej Sloveniji, zahtev po narečni uravnoteženosti baze ni bilo težko doseči. Manjši del govorcev so predstavljali študenti Mariborske univerze in njihovi svojci.

2.3 Oblikovanje vprašanj in vprašalnika

Vsak od govorcev je prejel pisna navodila za snemanje in izpolnil anketo za pridobitev podrobnejših informacij o njem. Anketa je bila vrnjena na univerzo v Mariboru. Najpomembnejša navodila za snemanje so bila ponovljena na vprašalniku, ki je bil narejen za vsakega govorca posebej na podlagi vrnjene ankete.

Med zaposlenimi na Pošti Slovenija je bilo izbranih 1450 posameznikov. Skupaj z drugimi kandidati je bilo uspešno končanih 1399 snemanj, od katerih se je izbralo 1000 klicev.

Teme, ki so povzročale govorcem največ problemov, so bile uvrščene na vprašalnik. Da bi se izognili zmedenosti govorcev, so bile brane besede in stavki grupirani.

2.4 Transkripcija

Transkripcija je bila narejena z orodjem za transkripcijo SpDatLabel, katero je bilo razvito na mariborski univerzi. Transkripcijski proces je potekal v dveh fazah: najprej so bile besede prepisane, nakar so prepisu bile dodane različne določbe.

Prepisovalo je pet različnih piscev; vsi so imeli univerzitetno izobrazbo. Delali so največ 4 ure (prepisali so povprečno osem govorcev) na dan, da bi se izognili napakam, ki nastajajo zaradi utrujenosti. Prepisovalci so uporabljali priročnik v slovenščini.

Izvršil se je avtomatski pregled napak pri črkovanju in napak v stavčni sintaksi. Dodatno je bilo izvršeno še dvojno preverjanje 2000 prepisov. To so bili predvsem prepisi, ki niso prišli skozi avtomatko preverjanje. 3,7% teh prepisov je bilo ocenjenih kot napačnih.

Slovarski termini so bili pisani z malimi črkami. Lastna imena, imena podjetij in geografska imena se začnejo z veliko začetnico, kakor je določeno s slovenskimi slovničnimi pravili.

Napačno izgovorjene besede, ki so bile vseeno razumljive, so bile označene z eno zvezdico, ki je bila pozicionirana na levi strani besede, katera je bila napačno izgovorjena (npr. *konj namesto napačno izgovorjene besede „kojn“).

Besede, ki imajo pred seboj postavljeno zvezdico, vsebujejo nepravilno izgovorjavo kot je izgovorjava z dodatnim ali izpuščenim zlogom. Zvezdica ni uporabljena, kadar gre za narečno besedo. V govoru je vsaka napačno izgovorjena beseda označena individualno.

Besede, ki so popolnoma nerazumljive, so označene z dvema zaporednima zvezdicama, ki sta ločeni od sosednjih besed s presledkom.

Besede, ki jih govorec ni končal, so upoštevane kot napačno izgovorjene besede in so označene z zvezdico na začetku besede.

Če je govor prekinjen zaradi snemalne napake, se pri zapisu uporabi naslednje simbole:

Začetek prekinjenega govora: ~prepis

Konec prekinjenega govora: prepis~

Začetek in konec prekinjenega govora: ~prepis~

Predpisane so štiri kategorije zvokov medtem, ko govorec ne govori. Zvoki so prepisani, če so točno določeni. Dogodek (zvok) je prepisan na mesto, kjer se je pojavil, z uporabo karakteriziranih simbolov v oglatih oklepajih. Šum, ki se pojavi preko ene ali več besed, je prepisan od začetka, preden je prekril prvo besedo.

Prvi dve kategoriji sta določeni za šume, ki izvirajo od govorca in drugi dve kategoriji za šume, ki izvirajo iz okolice. Zvoki, ki izvirajo od govorca ponavadi ne prekrivajo ciljnega teksta, zvoki, ki izvirajo iz okolice pa seveda lahko nastanejo istočasno z govorom.

[fil]: zapolnjen premor (Filled pause). Primeri: uh, um, er, ah, mm.

[spk]: šum govorca (Speaker noise). Vse vrste zvokov in šumov, ki nastanejo med govorjenjem govorca, ki niso del predpisanega teksta: cmokanje ustnic, kašljanje, tleskanje z jezikom, glasno dihanje, smeh, vzdihovanje.

[sta]: stalen šum (Stationary noise). V to kategorijo spadajo zvoki iz ozadja, ki niso prekinjeni in imajo več ali manj stabilno amplitudo spektra. Primer: zvok avtomobila, zvok s ceste, zvok mobilnega telefona, zvočne motnje na javnih mestih.

[int]: šum v presledkih (Intermittent noise). V to kategorijo spadajo šumi, ki se pojavijo samo enkrat (loputanje z vrati), ali je med njimi krajši premor (zvonjenje telefona) ali pa se jim skozi čas spremeni barva (glasba). Primeri: glasba, govor iz ozadja, jok otroka, zvonjenje telefona, hišni zvonec, šelestenje papirja.

Slovarski termini imajo samo en način črkovanja in so črkovani po slovenskih pravilih. Tuje besede so črkovane v tujem jeziku.

Številčna zaporedja (številke, čas, datum, denarni znesek) so črkovani.

Vse besede, ki so se pojavile v posneti bazi, so zbrane v fonetičnem leksikonu. Fonetične transkripcije besed v leksikonu so bile izdelane avtomatsko s konverterjem, razvitim na Univerzi v Mariboru, nato pa še ročno pregledane. Transkripcije so zapisane s SAMPA fonetično abecedo.

3 Vsebina baze

Slovenska baza SpeechDat vsebuje 43 izgovorjav vsakega govorca. Govor je deloma bran, deloma spontan. Dolžina posnetega govora vsakega govorca je približno šest minut. Opis vsebine klicev prikazuje Tabela 1.

Tabela 1: Vsebina klicev

Aplikacijske besede	6 aplikacijskih besed
Zaporedje samostojnih številk	1 zaporedje 10 samostojnih številk
Vezane številke	1 številka predloge 1 telefonska številka 1 številka kreditne kartice 1 identifikacijska koda
Datumi	1 spontan datum 1 predložen datum 1 relativna oz. splošna časovna fraza
Fraze za iskanje ključnih besed	1 fraza za iskanje ključnih besed z aplikacijsko besedo
Samostojne številke	1 samostojna številka
Črkovne besede/fraze	1 spontano (ime govorca) 1 ime kraja 1 umetna za ustrezno pokritost črk
Denarni znesek	1 denarni znesek
Naravna števila	1 naravno število
Imena za imenike	1 spontano (ime govorca) 1 spontano (mesto odraščanja) 1 ime mesta (nabor 500) 1 ime podjetja/ustanove (nabor 500) 1 ime in priimek (nabor 150)
Vprašanja	1 pretežno “da” vprašanje 1 pretežno ”ne” vprašanje
Stavki	9 fonetično bogatih stavkov
Časovne fraze	1 spontana fraza (trenuten čas) 1 časovna fraza (besedna oblika)
Besede	4 fonetično bogate besede

3.1 Aplikacijske besede

Tabela 2: Seznam aplikacijskih besed

Ključna beseda	Funkcija	Opis
osnovni IVR ukazi		
slovenski jezik	<language>	jezik storitve, npr. francoski (naravni jezik podatkovne baze)
končaj	<terminate>	Izhod iz aplikacije
menijske operacije		
meni	<menu>	Vrnitev v glavni meni
pomagaj, pomoč	<help>	Zahteva po informaciji ali možnostih v meniju, za trenutni dialog
upravljanje s funkcijami		
prekliči	<cancel>	Preklic trenutne operacije npr. klicanja, urejevanja, vnašanja
ustavi	<stop>	Ustavitev trenutne funkcije npr. predvajanja zvoka
nadaljuj	<continue>	Nadaljuj ustavljeno funkcijo ali začni obdelavo naslednjega elementa
ponovi	<repeat>	Ponovi zadnjo funkcijo ali ukaz, npr. predvajanje zvoka
funkcije za klicanje		
operater	<operator>	preveži klic na operaterja
kliči, klic	<call>	vzpostavitev telefonske povezave po imenu
izbiraj, izbor	<dial>	vzpostavitev telefonske povezave po številki
ponovno izbiraj	<redial>	Ponovno vzpostavi prejšnjo telefonsko povezavo
funkcije imenika		
imenik	<directory>	prehod na imenikov podmeni ali priklic imenikovih vnosov
navedi	<list>	priklic imen, sporočil, možnosti programiranja, itd.
predhodni	<previous>	pojdi nazaj za en vnos ali predvajaj predhodno sporočilo
naslednji	<next>	pojdi naprej za en vnos ali predvajaj naslednje sporočilo

konec	<end>	pojdi na zadnji vnos v seznamu ali na zadnje sporočilo
funkcije za upravljanje		
dodaj	<add>	dodaj ali vrini vnos, npr. ime in številka
menjaj	<change>	spremeni vnos, npr. ime in številko
zbriši, briši	<delete>	briši vnos ali sporočilo
shrani	<save>	shrani trenutni vnos ali sporočilo
funkcije sporočil		
predvajaj	<play>	predvajaj sporočilo ali datoteko
snemaj	<record>	posnemi sporočilo, glasovno pošto ali datoteko
pošlji	<send>	pošlji glasovno pošto
funkcije za programiranje		
programiraj, program	<program>	programiranje naprednih možnosti ali prehod v podmeni za programiranje

3.2 Zaporedje osamljenih številk

3.2.1 Posamezne številke

Vsak govorec prebere eno osamljeno številko, ki jo nenačrtno izbere od 0 do 9.

3.2.2 Niz števk

Z nizom števk se pridobi primere izgovorjav vseh števk za vsakega govorca. Vsak vprašalnik je bil sestavljen iz zaporedja števk od 0 do 9, ki je bilo naključno generirano, da bi se tako ognili predvidljivi ali monotoni izgovorjavi.

3.3 Vezane številke

3.3.1 Številka predloge

Prebrana številka predloge je sestavljena iz petih števk.

3.3.2 Telefonska številka

Telefonska številka je v obliki, kakršni se pojavlja v Sloveniji:

območna koda in številka

območna številka = 060x ali 06x ali 041

in

številka = xx-xxx ali xxx-xxx ali xxx-xx-xx

Primeri:

062/46-935

0609/97-825

0601-545-58-31

Območna koda 041 nam pove, da je to GSM številka.

3.3.3 Številka kreditne kartice

To je 15 – 16 števk. Številke, presledki in prezentacija številk prikazuje številke, katere so enake tistim, na kreditnih karticah. Uporabljena sta dva formata:

xxxx xxxx xxxx xxxx (16-mestna številka VISA/MasterCard/JCB/Discovers card)

xxxx xxxxxx xxxxx (15-mestna številka American Express)

Številke so izbrane iz stalne liste s 150 številkami, ki so posredovane SpeechDat partnerjem.

Vsaka druga številka iz tega seznama, je bila preoblikovana v 15-mestni American Express format.

3.3.4 Identifikacijska številka

Številka je sestavljena iz niza šestih števk, kar je podobno kot za številko predloge, vendar so te številke izbrane iz seznama 150 številk, ki so bile posredovane partnerjem SpeechDat.

3.4 Datumi

3.4.1 Spontan datum

Govorec je bil naprošen, da pove svoj datum rojstva.

3.4.2 Predložen datum

Številka je sestavljena iz dneva v tednu, dneva, meseca in leta.

Uporabljeni so bile naslednje besede:

dnevi v tednu = ponedeljek, torek, sreda, četrtek, petek, sobota, nedelja

dan = prvi, drugi, tretji, četrti, peti, šesti, sedmi, osmi, deveti, deseti, enajsti, dvanajsti, trinajsti, štirinajsti, petnajsti, šestnajsti, sedemnajsti, osemnajsti, devetnajsti, dvajseti, enaindvajseti, dvaindvajseti, triindvajseti, štiriindvajseti, petindvajseti, šestindvajseti, sedemindvajseti, osemindvajseti, devetindvajseti, trideseti, enaintrideseti

mesec = januar, februar, marec, april, maj, junij, julij, avgust, september, oktober, november, december

leto = številke med 1108 in 2910

Primer: nedelja, sedmi januar, 2316

3.4.3 Relativna oz. splošna časovna fraza

Uporabljene so bile naslednje časovne fraze:

danes
jutri
pojutrišnjem
včeraj
predvčerajšnjim
naslednji dan
prejšnji dan
dan po tem

dan pred tem

naslednji teden

prejšnji teden

naslednji mesec

prejšnji mesec

novo leto

velika noč

božič

poleti

spomladi

jeseni

pozimi

3.5 Črkovne besede/fraze

Slovenski jezik nima standardnih imen za črke, tako da govorci niso bili navajeni črkovati. Vsak govorec je črkoval tri besede: mesto, svoje ime (spontano) in niz črk. Bile so uporabljene črke iz slovenskega jezika. Skupaj se je pojavilo 23727 črk v črkovanih besedah. V tabeli 3 so napisane uporabljene črke, imena črk, ki so urejena po številu pogostosti uporabe in število uporabe. V splošnem sta prvo in drugo ime črk najbolj pogosta, okoli 99%.

Tabela 3: Pogostost črk

Črke	Imena	Število izgovorjav
a	A	1922
b	B BE BA BO EB	837
c	C CE CA CO EC	851
č	Č ČE	742
d	D DE DO DI	921
e	E	1084
f	F EF FE	792
g	G GE GO EG GA	774
h	H HA HE HI HO EH	743
i	I	1147
j	J JE EJ JA	1063

k	K KA KE KO	909
l	L EL LE LI	938
m	M EM AM ME	1018
n	N EN NE NO	1234
o	O	1003
p	P PE PA EP	762
r	R ER RA RE	1137
s	S ES SA SE SO	843
š	Š EŠ ŠE ŠO	780
t	T TE ET TA TI	981
u	U UJ	756
v	V VE VA VO EV	862
z	Z ZE ZA ZO ZU	810
ž	Ž ŽE ŽA	816

3.5.1 Spontano ime govorca

Vsak govorec je črkoval svoje ime.

3.5.2 Ime kraja

Govorec je črkoval ime mesta, ki je bilo eno izmed 100 največjih mest v Sloveniji.

3.5.3 Umetna za ustrezno pokritost črk

Govorec je črkoval niz črk, ki so bile izbrane tako, da niso bile že črkovane v prejšnji zahtevi, ko so črkovali ime kraja, da je bila pokritost črk večja.

3.6 Denarni znesek

Vsak govorec je prebral en denarni znesek. Bili sta izbrani dve obliki:

xxxx tolarjev

xx tolarjev x0 stotinov

kjer je x številka od 0 do 9. Bilo je poskrbljeno, da so se pomembni slovarski termini pojavili v zahtevanih številkah.

3.7 Naravna števila

Vsak govorec je prebral eno naravno število. Uporabljena so bila večja števila v obliki:

xxxxx

kjer je x številka od 0 do 9. Bilo je poskrbljeno, da so se pomembni slovarski termini pojavili v zahtevanih številkah.

3.8 Imena za imenike

3.8.1 Spontano ime (ime govorca)

Od govorca je bilo zahtevano, da pove svoje ime.

3.8.2 Spontano ime kraja (mesto odraščanja)

Govorec je bil vprašan po mestu odraščanja.

3.8.3 Ime mesta (nabot 500)

Ime mesta je bilo izbrano med imeni 500 največjih mest v Sloveniji.

3.8.4 1 ime podjetja/ustanove (nabor 500)

Ime podjetja oziroma ustanove je bilo izbrano iz nabora 500 podjetij, katera imajo največje število zaposlenih v Sloveniji.

3.8.5 Ime in priimek (nabor 150)

Izbrano je bilo ime in priimek iz nabora 150 najbolj pogostih imen in priimkov.

3.9 Da/Ne vprašanja

Za vprašanje s pretežnim odgovorom DA so bili govorci vprašani, če je Slovenščina njihov materni jezik.

Za vprašanje s pretežnim odgovorom NE so bili govorci vprašani, ali kličejo iz brezžičnega telefona.

3.10 Fonetično bogati stavki

Bilo je oblikovanih 9 področij s100 stavki. Stavki so bili vsebinsko različni. Vzeti so bili iz knjig, časopisov, itd. Del stavkov je bil vzet iz voznega reda vlakov. Stavki so bili sestavljeni pretežno iz 9 besed. Bili so avtomatsko izbrani iz večjega nabora. Tabela 4 prikazuje pogostost pojava fonemov.

Tabela 4: Pogostost pojava fonemov

Monofon	*Pogostost*
p	8187
b	3715
m	7990
f	522
v	5290
t	10685
d	5785
n	11916
s	10647
z	3839
r	9087
l	8118
k	6946
g	2534
e	9513
E	11056
a	22001
O	11820
i	17014
u	3453
o	1084
@	2687
w	4040
S	2788
Z	1092
tS	2927
ts	1484

j	10026
x	1994
	$\sum$ = 204538

3.11 Časovne fraze

3.11.1 Spontana fraza (trenuten čas)

Govorec je moral povedati trenuten čas.

3.11.2 Časovna fraza (besedna oblika)

Fraza je bila podana v analogni obliki naslednjega formata:

X Y Z

kjer je

X
četrt na
tri četrt na
pet minut čez
pol
točno
skoraj
deset do
danes
jutri

Y
ena
dve
tri
štiri
pet
šest
sedem
osem
devet
deset
enajst
dvanajst

Z
zjutraj
popoldan
dopoldan
ponoči
opoldne
opolnoči

Uporabljene so bile samo logične povezave.

Primer: četrt na ena zjutraj

3.12 Fonetično bogate besede

Vsak govorec je prebral štiri besede iz nabora 1500 slovenskih besed. Besede so bile izbrane iz različnih tekstov in slovarjev, izbor je bil avtomatski.

Tabela 5: Pogostost pojava monofona v fonetično bogatih stavkih

Monofon	*Pogostost*
p	565
b	480
m	564
f	122
v	496
t	2549
d	565
n	1428
s	1257
z	824
r	1196
l	917
k	906
g	349
e	779
E	848
a	3153
O	1108
i	3201
u	446
o	507
@	670
w	382
S	357
Z	275
tS	513
ts	179

j	551
x	223
	∑ = 25410

4 SpDatLabel - snemalno orodje za (Speech Databases) podatkovne zbirke

Postopek transkripcije je v razvoju govorne podatkovne baze eden najpomembnejših in tudi časovno najzahtevnejših procesov. S pomočjo učinkovitih transkripcijskih oz. snemalnih orodij lahko privarčujemo mnogo človeškega truda in strojne opreme, zato ta orodja igrajo vitalno vlogo.

Orodje SpDatLabel je bilo razvito za transkripcijo slovenske podatkovne baze SpeechDat, ki je bila posneta v projektu SpeechDat II (LE2-4001), lahko pa ga uporabimo tudi za druge podatkovne baze, ki ustrezajo formatu baze SpeechDat. Glavni cilj pri načrtovanju orodja SpDatLabel je bil čim bolj zmanjšati čas potreben za transkripcijo, pri čemer pa naj bo število transkripcijskih napak minimalno. Glede na ti dve glavni zahtevi, je bil SpDatLabel zamišljen kot orodje, ki naj čimbolj avtomatizira postopek transkripcije. Glede na to, da je v podatkovni bazi SpeechDat uporabljena ortografska transkripcija, je bilo mogoče doseči veliko stopnjo avtomatizacije.

Interaktivni del orodja SpDatLabel predstavlja grafični uporabniški vmesnik, sestavljen iz enega okna in naslednjih razdelkov:

> ***Podatki o govorcu.*** V tem razdelku se vnašajo podatki o posameznem govorcu, kot so številka snemanja, identifikacijska (ID) številka, spol, starost, itd. Večino teh podatkov je moč izbrati s pomočjo izvlečnih menujev, ki že vsebujejo določen nabor vnaprej določenih vrednosti.
>
> ***Podatki o zapisu in izgovorjavi.*** Ta razdelek vsebuje polja s kodo, besedilom in transkripcijo izgovorjave, pri čemer mora transkripcijo vnesti operater.
>
> ***Signalno okno.*** V signalnem oknu je prikazana izgovorjava, za katero se izvaja transkripcija. Operater ročno označi začetek in konec izgovorjave.

Postopek transkripcije se opravlja na govorčevi osnovi. Operater najprej vnese podatke o govorcu, ki so zbrani s pomočjo anketiranja govorca. Dejanska transkripcija se prične z izbiro kode prve izgovorjave. Prikažeta se predloženo besedilo in najverjetnejša

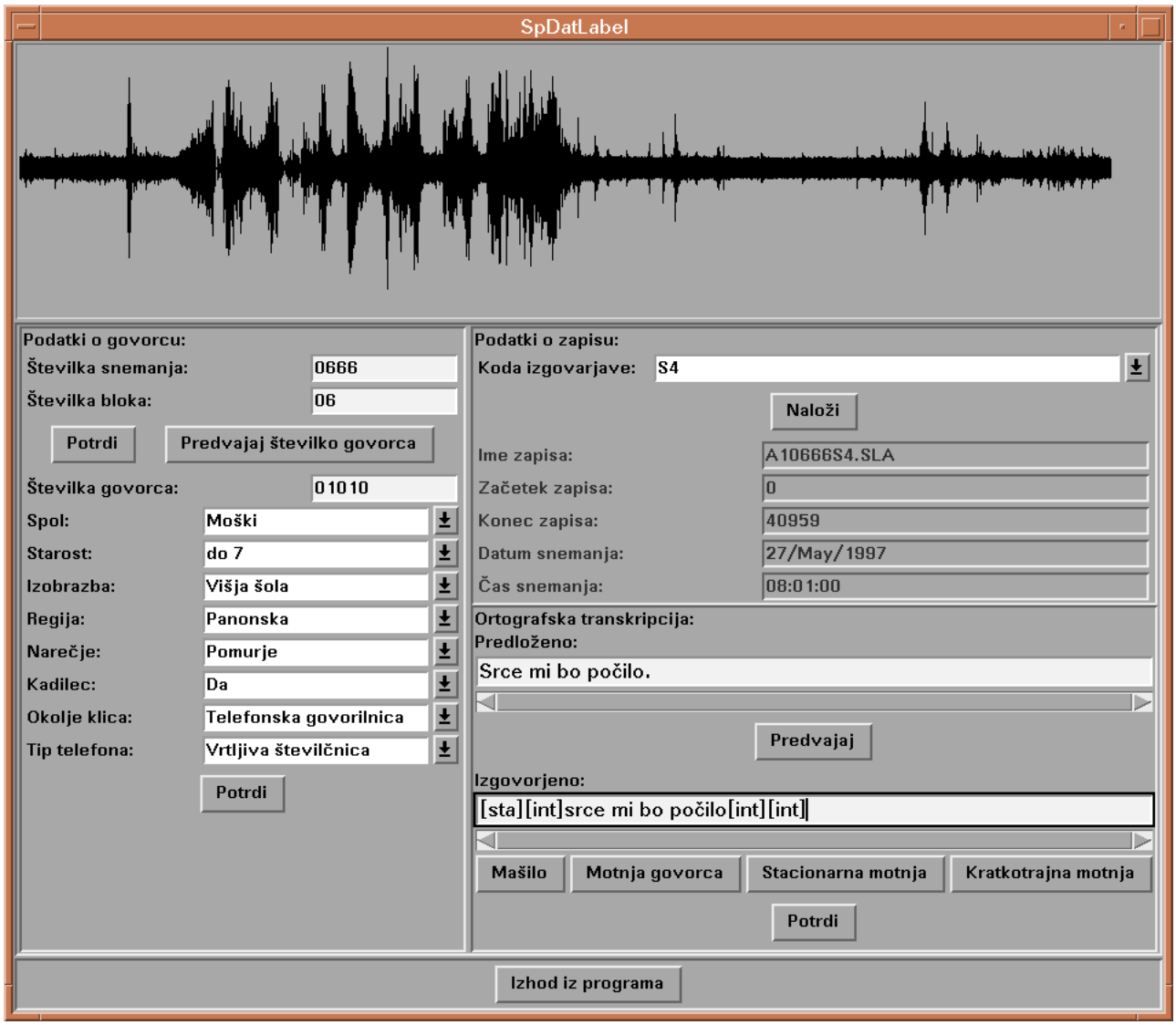
transkripcija. Predlagana transkripcija je izpeljana iz predloženega besedila z uporabo transkripcijskih pravil, definiranih v projektu SpeechDat. Operater nato posluša izgovorjavo in popravi predlagano transkripcijo. Ta postopek se izvede v dveh korakih. V prvem se izvede transkripcija govora, v drugem koraku pa so dodane oznake za posebni govor in ne govorne dogodke. V signalnem oknu se označijo meje izgovorjave. Ko se potrdi transkripcija prve izgovorjave, se samodejno naloži naslednja izgovorjava.

Orodje SpDatLabel je implementirano z uporabo programskega jezika Tcl/Tk/Tix. Tcl je skriptni jezik in Tk/Tix sta razširitvi jezika Tcl za razvoj grafičnega vmesnika. Zaradi dejstva, da so programi, napisani v Tcl/Tk/Tix relativno počasni, so nekateri, časovno kritični procesi implementirani v jeziku C/C++. Trenutno je SpDatLabel implementiran pod operacijskim sistemom HP-UX, vendar je enostavno prenosljiv na druge platforme. Med razširjeno uporabo orodja pri transkripciji slovenske podatkovne baze SpeechDat so bile poudarjene naslednje ugotovitve:

Ker je postopek transkripcije avtomatiziran, se operater lahko skoncentrira le na samo transkripcijo in mu ni potrebno skrbeti za pravilnost izvedbe postopka.

Postopek transkripcije se izvaja hitro in brez zapletov.

Ugotovljeno je bilo, da je pogostost pojavljanja napak relativno nizka. To sledi iz dejstva, da se najbolj verjetna transkripcija izvede avtomatsko, nato pa jo ročno popravi operater.

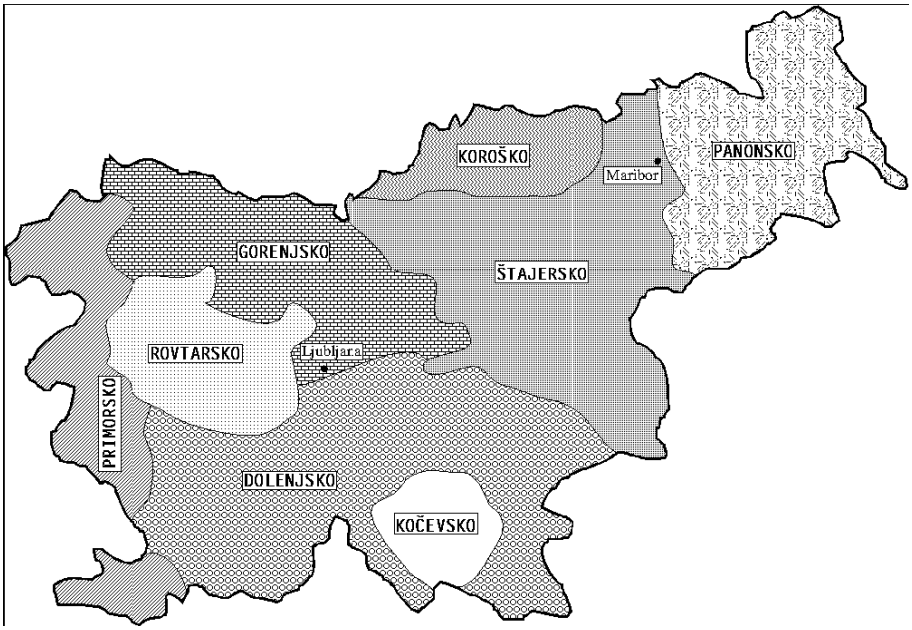


Slika 1: Orodje SpDatLabel

5 Demografske lastnosti govorcev

5.1 Narečje in regija

Slovenija je republika s približno 2 miliona prebivalci. Po podatkih, ki so zapisani v “Karti slovenskih narečij” iz leta 1993, je Slovenija razdeljena na 10 narečnih skupin.



Slika 2: Karta Slovenije z narečnimi skupinami

Narečno področje govorcev ima močan vpliv na njihov govor. Pomembni pa so tudi drugi dejavniki, kot je npr. jezikovno ozadje staršev ter kraj, kjer je govorec obiskoval šolo. Slovenija je razdeljena na deset narečnih področij. Število govorcev iz posameznega področja je bilo določeno glede na število prebivalcev v danem področju. Tabela 6 prikazuje porazdelitev govorcev po narečjih v bazi.

Tabela 6: Razporeditev govorcev po narečnih področjih

	Narečno področje	**Oznaka v bazi podatkov**	**Število govorcev**	**Odstotek govorcev (%)**
1	Panonsko	PAN	52	5.24
2	Štajersko	STA	272	27.39
3	Koroško	KOR	42	4.23
4	Dolenjsko	DOL	147	14.80
5	Kočevsko	KOC	15	1.51
6	Rovtarsko	ROV	31	3.12

7	Gorenjsko	GOR	108	10.88
8	Primorsko	PRI	71	7.15
9	Maribor	MAR	86	8.66
10	Ljubljana	LJU	169	17.02

5.2 Karakteristike govorcev

5.2.1 Spol in starost

Spol govorca ima izreden vpliv na kvaliteto govora in pogojuje vsaj naslednje štiri dejavnike:

- višina in jakost – ženske govorijo v povprečju z višjim tonom kot moški;
- spektralna ovojnica – ženske imajo bolj dihnjen glas;
- pravilnost izgovorjave – ženske se bolj držijo standardne izgovorjave kot moški;
- slovar in skladnja – opažene so bile razlike med spoloma pri uporabi besed in skladnji, ki pa so opazne le pri spontanem govoru.

Starost vpliva vsaj na dva dejavnika govora:

- kvaliteta govora – vpliv starosti na avtomatsko razpoznavanje govora ni znan; ker pa so ljudje sposobni določiti starost govorca iz njegovega govora so se pri projektu SpeechDat projekta odločili razlikovati govorce glede na njihovo starost;
- slovar in skladnja – opažene so podobne razlike kot pri spolu govorcev.

Tabela 7: Razporeditev govorcev glede na starost in spol

Starostne skupine	**Oznaka v bazi podatkov**	**Število moških govorcev**	**Število ženskih govork**	**Skupno (%)**
Pod 16	12	1	5	0.60
16-30	23	157	176	33.53
31-45	38	159	238	39.98
46-60	53	140	116	25.78
Preko 60	78	1	0	0.10

V večini kultur je zelo težko razlikovati med socialnimi razredi in določiti vpliv le-teh na govor njihovih pripadnikov. V Sloveniji ima verjetno največji vpliv dosežena izobrazba govorca, zato so se odločili razlikovati govorce po tem kriteriju. Porazdelitev govorcev po izobrazbi prikazuje Tabela 8.

Tabela 8: Razporeditev govorcev glede na stopnjo izobrazbe

Stopnja izobrazbe	Oznaka v bazi podatkov	Število govorcev	Skupno (%)
Osnovna šola	2	39	3.93
Poklicna šola	3	102	10.27
Srednja šola	4	269	27.09
Višja šola	5	447	45.02
Visoka šola	6	100	10.07
Magisterij	7	35	3.52
Doktorat	8	1	0.10

Tabela 9: Razporeditev govorcev glede na kadilske navade

Kadilske navade	Oznaka v bazi podatkov	Število govorcev	Skupno (%)
Nekadilec	NONSMOKER	729	73.41
Kadilec	SMOKER	264	26.59

5.3 Okolje

Pomemben faktor pri snemanju govora je okolje, v katerem se je nahajal govorec med snemanjem. Ker je namen SpeechDat baze čim bolje odraziti realne razmere, so lahko bili klici opravljeni iz poljubnega okolja.

Tabela 10: Okolje v katerem se govorec nahaja

Okolje	Oznaka v bazi podatkov	Število klicev	%
Dom	HOME	728	72.80
Pisarna	OFFICE	211	21.10
Javni prostor	PUBLIC	27	2.70
Telefonska govorilnica	BOOTH	22	2.20

Tovarna	FACTORY	1	0.10
Drugo	OTHER	11	1.10

6 SpeechDat in projekti

Podatkovne baze za ustvarjanje glasovnih gonilnih telefonskih servisov. Projekt CEC LE2-4001 zajema:

- Aplikacijo - besede in izražanja
- Širino sekvenc
- Finančni izračun
- Datum in čas izražanj
- Črkovanje
- Direktorij - pomoč – podpora besedišču
- Da / Ne odgovori
- Bogatišče fonetičnih stavkov in besed

6.1 Projekti v teku

6.1.1 OrienTel

Več jezikovni dostop do interaktivnih komunikacijskih servisov za Sredozemlje in Srednji vzhod.

Razvoj jezikovnih virov osnovanih za govorno telefonijo nameščenih med področji Maroka in zalivskih držav, vključno z različnimi oblikami nemščine, francoščine, angleščine, arabščine, grščine, turščine in hebrejščine. OrienTel upravlja akademski in gospodarskih konzorcij, ki ga sestavljajo:

- ScanSoft
- IBM
- University of Patras
- Natural Speech Communication
- Lucent Technologies

- ELDA
- Knowledge
- Siemens
- UPC

6.1.2 SpeechDat-Car

Devet prenosnih in mobilnih telefonskih omrežnih podatkovnih baz, vsak obsega 600 posnetkov.

6.1.3 SALA

SpeechDat projekt, ki obsega celotno Latinsko Ameriko.

6.1.4 SpeechDat(E)

SpeechDat projekt, ki se izvaja v Vzhodni Evropi.

Ta projekt je osredotočen na govorne jezikovne vire, kot so govorne in e podatkovne baze za fiksna – stacionarna omrežja vključno z povezavo s slovarjem knjižnega jezika.

Te podatkovne zbirke se uporabljajo za izobraževanje in preizkušanje tipičnih vsakdanjih telefonskih servisov, kot fonetično obogateno gradivo, ki se uporablja za šolanje bolj naprednih, slovnično neodvisnih sistemih za govorno prepoznavanje.

Obsegajo zbirko najbolj branih in govorjenih besed (različne vrste izgovorjav).

Podatkovna zbirka je oblikovana na 1000 – 2500 govorcih, izenačenih po spolu, starosti in prepoznavnosti narečja. Namen projekta je izdelati podatkovno zbirko jezikov, kot so: Ruski (z 2500 govorci), Češki, Slovaški, Poljski in Madžarski jeziki (z 1000 govorci). Ta podatkovna zbirka bo služila kot pomemben vir pri predstavitvi

glasovnih gonilnih telefonskih servisov in uporabnih izvedb.

6.2 Končani projekti

6.2.1 SpeechDat(II)

25 stacionarnih in mobilnih telefonskih omrežij, vsak od njih ima 500-5000 govorcev: 3 potrjene govorne podatkovne zbirke

6.2.2 SpeechDat(M)

8 stacionarnih telefonskih omrežij, 1000 govorcev vsak; 1 mobilna telefonska mreža, 300 govorcev.

6.3 Sorodni projekti

6.3.1 SPEECON

Govorne podatkovne zbirke za potrošniško uporabo.

6.3.2 LILA

Zbiranje Jezikovnih virov v Aziji. Mobilna in fiksna – stacionarna telefonija, razvrstitev snemanj v Koreji, Indiji vključno z Angleščino, Kitajska vključno z Tajsko, Malezijo, Filipini, Japonsko, Indonezijo in Tajsko.

Cilj projekta LILA je zbiranje velikega števila govornih podatkovnih baz za urjenje sistemov za samodejno prepoznavanje govora v Azijsko Pacifiškem območju. Govorne podatkovne baze so zbrane preko mobilne telefonske mreže. LILA konzorcij je sestavljen iz velikega števila gospodarskih družb. Vsaka družba je vodilna pri produkciji podatkovnih zbirk. Konzorcij deli podatkovne zbirke proizvedene v projektu. En cilj projekta naj bi bil dosežen v letu 2006.

LILA projekt je sklad gospodarskega konzorcija. Konzorcij je dostopen članom iz gospodarstva ali članom javnega značaja. Vsi člani imajo dostop do podatkovne zbirke izdelane znotraj konzorcija. Univerza Politehnike v Kataloniji (UPC) (Španija) koordinira projekt. Pravnomočnost podatkovnih zbirk izvršuje holandski inštitut SPEX.

Območje obsega veliko količino jezikov in dialektov govorjenih na področjih kot so Avstralija, Kitajska, Indija vključujoč angleški jezik, Indonezija, Japonska, Koreja, Malezija, Nova Zelandija, Filipini, Tajska, Taiwan, Vietnam… Za vsako področje je bil narejen vpogled v prebivalstvo, skozi gospodinjstva, uradne jezike, govorjene dialekte, sprejemljiva števila govorcev na jezik (in/ali dialekt) in razpoložljivost podatkov (stacionarni, mobilni, in-car, potrošni).

Znotraj družine projektov SpeechDat (SpeechDat, SpeechDat car, SALA, Speecon), so govorne podatkovne baze previdno določene glede na vsebino (uporaba besed in fraz, fonetično bogate besede in stavki), delovanje glede na spol, regije z dialekti, starost

govorcev, število govorcev in snemalne naprave. V projektu je bilo oblikovanih nekaj razširjenih oblik in tonskih posnetkov za usmeritev. Osnovan je bil pravnomočen oris za garancijo kvalitete podatkovnih zbirk.

Trenutno posnete podatkovne zbirke so prikazane v tabeli 11.

Tabela 11: Trenutno posnete zbirke

Jezik	Področje	Govorci	Družba
Korejski	Koreja	1000	NSC
Hindujščina (kot 1 jezik)	Indija	2000	Siemens
Hindujščina (kot 2 jezik)	Indija	1800	Motorolla
Indijska Angleščina	Indija	1800	Nuance
Kitajska Mandarinščina	Kitajska	1800	Microsoft

7 Zaključek

SpeechDat II je skupen projekt 25 različnih inštitucij in ustanov iz 15 evropskih držav. Namen projekta je bil izdelati govorne baze podatkov, namenjene za razvoj avtomatskih telefonskih sistemov govornega dialoga in določitev standardov ter priporočil za izdelavo baz podatkov in njihovo vrednotenje. Za slovenski jezik (Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru) je bila izdelana baza 1000 govorcev, ki je bila posneta z avtomatskim strežnikom, vezanim na ISDN telefonsko omrežje. Pri izbiri govorcev so bile upoštevane zahteve za uravnoteženost po spolu, starosti in narečju govorcev. Slovenska baza vsebuje 43 izgovorjav vsakega od govorcev v skupni dolžini okoli pet minut. Baza je že bila uporabljena pri razvoju govornih aplikacij na Fakulteti za elektrotehniko, računalništvo in informatiko.

SpeechDat slovenska baza izgovorjav za stacionarno telefonijo je bila dokončana Decembra 1997 in nato uspešno validirana v nizozemskem podjetju SPEX. Shranjena je na petih CD-ROM ploščah, od katerih vsaka vsebuje posnetke 200 govorcev. Posnetki so shranjeni v A-law, 8 bitnem, 8 kHz formatu. Baza je dosegljiva preko ELRA.

8 Viri

www.feri.uni-mb.si

www.dsplab.uni-mb.si

www.elda.org

www.speechdat.org

www.elda.org/catalogue/en/speech/design/s0056desc.doc