Jerneja Žganec Gros[1], France Mihelič[2], Simon Dobrišek[2] UDK811.163.6-25:004.5
[1]*Alpineon razvoj in raziskave d.o.o., Ljubljana*
[2]*Univerza v Ljubljani, Fakulteta za elektrotehniko in računalništvo*
jerneja@alpineon.com, mihelicf@fe.uni-lj.si, simond@luks.fe.uni-lj.si,

GOVORNE TEHNOLOGIJE: PRIDOBIVANJE IN PREGLED GOVORNIH ZBIRK ZA SLOVENSKI JEZIK

Govorne zbirke so nepogrešljive pri raziskovalnem delu na področju tehnologij govorjenega jezika. Vsebujejo predvsem računalniško berljive posnetke govora. Posnetkom so vedno priloženi še podatki, ki na različne načine opisujejo posneti govor. Priloženi podatki so opisi govornih dejavnikov, dejavnikov govorcev ter zapisi in označitve posnetega govora. Pogosto govorni zbirki priložimo še navodila za uporabo ter rezultate izbranih analiz govornih posnetkov. Članek opisuje osnovne gradnike govornih zbirk ter postopek zasnove, snemanja, segmentacije, označevanja in analize govornih zbirk. V sklepnem delu se nahaja pregled govornih zbirk za slovenski jezik.

1 Uvod

Govorne tehnologije, predvsem to velja za sintezo in prepoznavanje govora, nezadržno prodirajo v naše življenje. Na tržišču se je v zadnjih letih pojavilo kar nekaj solidnih sintetizatorjev in prepoznavalnikov govora, nekateri med njimi podpirajo tudi slovenski jezik. Uporabljamo jih v samodejnih informacijskih centrih, v govornih portalih, za glasovno prebiranje elektronske pošte ipd.

Razvoj in raziskave s področja govornih tehnologij za slovenski jezik se izvajajo na Fakulteti za elektrotehniko Univerze v Ljubljani (Luks), na Fakulteti za elektrotehniko, računalništvo in informatiko na Univerzi v Mariboru (FERI), na Fakulteti za računalništvo in informatiko Univerze v Ljubljani (FRI), na Inštitutu Jožef Stefan (IJS), na Naravnoslovno tehniški fakulteti Univerze v Ljubljani ter v podjetjih Masterpoint d. o. o. (Masterpoint), Hermes Softlab d. d. (HSL) in Alpineon razvoj in raziskave d. o. o. (Alpineon).

Razvoj govornih tehnologij, predvsem sinteze in prepoznavanja govora, ni pogojen le s tehnologijo izgovora, temveč tudi z izvajanjem osnovnih raziskav govora in jezika. Vse več govornih zbirk in rezultatov jezikovnih študij je dostopnih tudi v našem prostoru. Govorne zbirke so nepogrešljive pri raziskovalnem delu na področju govornih tehnologij. Predstavljajo pomemben člen osnovne infrastrukture za razvoj govornih tehnologij za posamezno jezikovno področje (Kačič 1998).

Govorne zbirke vsebujejo predvsem računalniško berljive posnetke govora (Gibbon 1997; Dobrišek 2001). Posnetkom so vedno priloženi še podatki, ki na različne načine opisujejo posneti govor. Priloženi podatki so opisi govornih dejavnikov, dejavnikov govorcev ter zapisi in označitve posnetega govora.

V članku opisujemo osnovne gradnike govornih zbirk. Zasnovo govorne zbirke lahko logično razstavimo v tri zaporedne korake, ki jih podrobneje predstavljamo v nadaljevanju članka. Prvi korak predstavlja izbira besedila, potrebnega za snemanje govorne zbirke, oz. izbira govorne situacije v primeru, ko gre za snemanje spontanega govora. Drugi korak predstavlja snemanje govornega gradiva. Sledi zadnji korak, ki ga sestavljajo segmentacija, označevanje in analiza govornega gradiva.

Članek sklenemo s pregledom govornih zbirk za slovenski jezik, zbranih v laboratorijih, ki na našem prostoru delujejo na področju govornih tehnologij.

2 Govorne zbirke

Govorne zbirke, pravimo jim tudi zbirke govorjenega jezika, so nepogrešljive pri raziskovalnem delu na področju tehnologij govorjenega jezika. Vsebujejo predvsem računalniško berljive posnetke govora (Gibbon 1997; Dobrišek 2001). Posnetkom so vedno priloženi še podatki, ki na različne načine opisujejo posneti govor. Priloženi podatki so opisi govornih dejavnikov (npr. spol, starost, poklic, narodnostno-narečno območje), dejavnikov govorcev (npr. govorčevo trenutno razpoloženje, zvočne lastnosti okolja ob snemanju) ter zapisi in označitve posnetega govora. Pogosto so priložena še navodila za uporabo zbirke ter rezultati izbranih analiz govornih posnetkov (Fourcin 1989).

Govorne zbirke navadno vsebujejo samo nekatere od naštetih podatkov. Govorni posnetki pa so vedno opremljeni vsaj z nekaterimi dodatnimi podatki. Kateri od naštetih podatkov so dejansko vključeni v zbirko, je odvisno od namena uporabe zbirke (Dobrišek 2001).

Govorne zbirke moramo razlikovati od drugih podobnih zbirk, povezanih z jezikom in govorom. Tako govorne posnetke, ki niso računalniško berljivi ali niso opremljeni s primernimi dodatnimi podatki, ne štejemo za prave govorne zbirke. Pri tem mislimo na radijske in televizijske govorne posnetke, shranjene na že zastarele zvočne medije. S prenosom takih posnetkov na računalniško berljiv medij in s pripravo potrebnih dodatnih podatkov bi iz teh posnetkov lahko nastale pomembne govorne zbirke.

Po drugi strani se govorne zbirke razlikujejo tudi od besedilnih zbirk, ki jih pridobivamo za potrebe jezikovnega modeliranja, jezikoslovnih analiz naravnega jezika in za gradnjo samodejnih sistemov za prevajanje. V te zbirke so vključena večinoma le pisana besedila, pridobljena iz različnih virov (romani, časopisi, gledališka dela ipd.). Te zbirke praviloma ne vključujejo govornih posnetkov, kvečjemu ortografsko prepisana govorjena besedila. Največji tovrstni zbirki za slovenski jezik sta FIDA in Nova beseda.

3 Govorni posnetki

Govorni posnetki so najpomembnejši del govornih zbirk. Sem uvrščamo vse posnetke, ki jih lahko pridobimo v povezavi z delovanjem človeških govoril med govorjenjem. Najpogostejši so mikrofonski posnetki zvočnih signalov, ki se pojavljajo med govorjenjem, in laringografski posnetki vibriranja glasilk v grlu. Uporabljajo pa se tudi vizualni in rentgenski posnetki govoril. Prvi se nanašajo zgolj na slikovne posnetke obraza, predvsem ustnic, drugi pa na globinske slike prereza celotnega sistema govoril. V zadnjem času se pogosto analizirajo tudi posnetki biomedicinskih signalov, ki se pojavljajo v govorilih in tudi v človeških možganih. Ti signali se pridobivajo med govorjenjem s posebno medicinsko opremo (elektromiografi, elektroencefalografi).

Večina raziskav na področju govornih tehnologij se še vedno nanaša predvsem na obdelavo posnetkov govornih zvočnih signalov, pridobljenih s pomočjo mikrofona (Dobrišek 2001). Tem signalom pravimo govorni signali. Vsekakor imajo tudi preostali omenjeni govorni posnetki poseben pomen, saj njihova analiza omogoča bistveno širši vpogled v delovanje človeških govoril in s tem tudi globlje razumevanje človeškega govora.

Govorne posnetke shranjujemo na raznovrstne magnetne, optične in druge sodobne pomnilniške medije. Pri tem je pomembno, da je izbrani medij računalniško berljiv. Praviloma se govorni posnetki shranjujejo kar v obliki računalniških datotek na lasersko berljive plošče CD-ROM in DVD-ROM.

4 Opisi govornih posnetkov

Govorna zbirka vsebuje tudi opise različnih dejavnikov, ki se nanašajo na govorne posnetke. Pri tem mislimo predvsem na govorne dejavnike in dejavnike govorcev, katerih govor smo snemali. Oboje je neposredno povezano z govorci, ki so sodelovali pri snemanju govora (Dobrišek 2001).

Govorci

Najbolj preproste govorne zbirke so zbirke z nekaj deset govorci (Gibbon 1997). Število govorcev znotraj tega razpona že omogoča statistično ustreznost zbirke. To zagotovimo z upoštevanjem standardnih statističnih postopkov pri izbiri vzorca populacije vseh govorcev (Marascuilo 1988).

Statistično ustreznost zbirke nenazadnje določa tudi njen končni namen. Pogosto se namreč že vnaprej omejimo le na ožje področje govorjenega jezika in le na ožjo populacijo govorcev. Na primer: gradnjo kakovostnih in od govorcev neodvisnih sistemov za prepoznavanje govora omogočajo govorne zbirke z nekaj sto govorci. Pri razvoju sistemov za samodejno tvorjenje govora pa se uporabljajo zbirke z le nekaj govorci. Pogosto je v takšne zbirke vključen en sam izkušen govorec, ki predstavlja učni primer in primerjavo bodočemu sintetizatorju govora (Gros 2000). Pri

tem je pomembno, da je govorec vedno na voljo zaradi morebitnih popravkov ter dosnemavanja govorne zbirke.

Dejavniki govorcev

Pri izbiri govorcev moramo upoštevati nekatere značilnosti, ki so povezane z njihovim govorom. Te značilnosti delimo na prehodne in trajne (Dobrišek 2001). Prehodne značilnosti so morebitne psihološke in fiziološke motnje (npr. počutje, bolezen, psihično stanje).

Med trajne značilnosti pa štejemo fiziološke in anatomske značilnosti (npr. spol, starost, težo, okvare na govorilih, kadilske in pivske navade) in geografske in socialno-jezikovne vplive na govor (izobrazbo in poklic govorcev, narodnostno-narečno področje trenutnega in morebitnega predhodnega bivališča govorcev, narodnostno-narečno področje govorčevih staršev).

Govorni dejavniki

Med govorne dejavnike uvrščamo prozodijske značilnosti govora, ki se nanašajo na trenutno razpoloženje govorca ter njegov pristop k tvorjenju govornih posnetkov (Dobrišek 2001). Posneti govor lahko tako označimo za hiter, počasen, napet, sproščen, odrezav, natančen, površen ipd.

4.1 Simbolni zapisi govora

Govorni signal je psevdonaključen, krajevno stacionaren signal, ki nosi informacijo. Govorni signal vsebuje več informacij, kot jih je v besedilu (npr. čustva, odnos govorca do teme in podobno).

S preučevanjem govornega signala s stališča izgovarjave ali slušnosti se ukvarja fonetika (Toporišič 1992; Srebot–Rejec 1988). Preučuje zvočno podobo jezika od glasu, naglasa v besedi do celotnega besedila. Govorimo lahko o fonetiki glasov, prozodike, zlogov, morfemov, besednih zvez, stavkov, povedi in besedila.

Za opisovanje fenomenov govorjenega besedila se uporabljajo predvsem 3 nivoji anotacij ali prepisov govorjenega besedila: grafemski prepis, fonetični prepis in prozodijske oznake.

Oznakovni del govorne zbirke, predvsem njen grafemski prepis, je pogosto zapisan po priporočilih TEI P3 (Text Encoding Initiative). Priporočila TEI P3 določajo konkretne oznake ISO standarda SGML (Standard Generalized Markup Language) in strukturo oznak SGML za raznovrstna besedila (Erjavec 1998).

Za označevanje govora na ožji fonetični ravni se je v svetu uveljavila mednarodna fonetična abeceda IPA (International Phonetic Alphabet) oz. njena računalniško berljiva poenostavljena preslikava MRPA (Machine Readable Phonetic Alphabet).

Fonetično abecedo MRPA za slovenske glasoslovne oznake je v medsebojnem sodelovanju izdelalo več slovenskih raziskovalnih institucij (Zemljak 2002).

Za označevanje prozodijskih fenomenov v govorjenem jeziku se uporabljajo posebne prozodijske anotacije (Mihelič 2000; Stergar 2000).

Fonetični prepis govora

Grafemski oziroma črkovni zapis govorjenega gradiva predstavlja osnovni simbolni zapis posnetega govora. Fonetični zapis govora uporablja osnovne enote, ki so povzete iz glasoslovja. To so fonemi ali alofoni. Tak simbolni zapis govora vsebuje več informacije od grafemskega.

Različna izbira anotacijskih enot predstavlja tudi različne možnosti zapisa ali prepisa govorjenega jezika. Poleg običajnega grafemskega zapisa tako govorimo o fonemskem, ožjem in širšem fonetičnem ter akustičnofonetičnem zapisu. Primeri različnih zapisov za eno poved so zbrani v preglednici 1.

način zapisa	primer zapisa
grafemski zapis	Krpan se ni dal zmesti.
širši fonetični zapis	k@rpan sE ni daU zmEsti .
ožji fonetični zapis	k@rpa:n sE ni da:U zmE:sti .
akustičnofonetični zapis	-k@r-pa:n sE ni: =daU zmE:s-ti .

Preglednica 1: *Različni načini zapisov ali prepisov besedila. Fonetični prepisi besed so zapisani s simboli slovenske MRPA abecede. [-] in [=] sta oznaki za nezvenečo in zvenečo zaporo pri akustičnofonetičnem zapisu fonemov zapornikov <p>,<k>,<t>,<d>,<g> in <b>.*

4.2 Označitve govornih posnetkov

Simbolni zapisi in prepisi posnetega govora večinoma zadoščajo pri raziskovalnem delu na področju tehnologij govorjenega jezika. To velja predvsem za tehniško modeliranje govora (Dobrišek 2001). Podrobnejša analiza govora pa zahteva natančnejše označevanje posnetkov govora.

Označevanje govornih posnetkov je postopek ročnega ali samodejnega določanja odsekov posnetih govornih signalov, ki jih obravnavamo kot akustične enote govorjenega jezika. Vsakemu odseku pripišemo simbolno oznako, ki v govornem signalu predstavlja oziroma označuje akustično enoto.

Glasovi so osnovne akustične enote govorjenega jezika. Govorne posnetke zato pogosto označujemo tako, da jih razčlenimo na zaporedje odsekov, ki predstavljajo fone fonemov ali alofonov. Tudi takšno označevanje je lahko bolj ali manj natančno. Natančnost je odvisna od števila upoštevanih glasovnih različic oziroma alofo-

nov ter morebitnega označevanja akustičnih dogodkov, manjših od samih fonov. Pri slednjih obravnavamo predvsem dogodke, povezane z govorilnimi organi, kot so tlesk, zapora, odpora, pripora, pridih.

Pri označevanju prav tako govorimo o fonemski, širši in ožji fonetični, akustičnofonetični ter prozodijski označitvi govornih posnetkov, podobno kot pri simbolnih zapisih in prepisih govora.

4.3 Opisi analiz in navodila

Govornim zbirkam pogosto priložimo raznovrstne rezultate analiz zbirke. To so navadno rezultati statističnih analiz, kot so frekvence (število) posameznih akustičnih enot ter sklopov akustičnih enot.

Govorno zbirko opremimo še z navodili za uporabo. To so podatki o strukturi zbirke, datotečnem sistemu in formatih računalniških zapisov. Za tehniško modeliranje govora je pomembno navesti še podatek o tem, kateri del zbirke je namenjen učnemu postopku, s katerim določamo parametre modelov akustičnih govornih enot, in kateri del zbirke je namenjen preizkušanju in vrednotenju teh modelov.

5 Postopek pridobivanja govorne zbirke

Postopek pridobivanja govorne zbirke lahko logično razstavimo v tri zaporedne korake, ki jih podrobneje predstavljamo v tem poglavju. Prvi korak predstavlja izbira besedila, potrebnega za snemanje govorne zbirke, oz. izbira govorne situacije v primeru, ko gre za snemanje spontanega govora. Drugi korak predstavlja snemanje govornega gradiva. Sledi zadnji korak – segmentacija, označevanje in analiza govornega gradiva.

5.1 Načrtovanje vsebine zbirke

Pri načrtovanju govornih zbirk velja, da naj bodo čim bolj obsežne. Žal si neomejeno velikih količin podatkov v zbirki ne moremo privoščiti. Zato moramo pazljivo načrtovati vsebino zbirke, da ta čim bolje predstavlja celotno izbrano področje govorjenega jezika (Gibbon 1997).

Ob načrtovanju govorne zbirke želimo doseči statistično ustrezno vzorčenje izbranega področja govorjenega jezika. Tovrstne ustreznosti govorne zbirke ne moremo zagotoviti z zgolj inženirskim pristopom, ker pri tem potrebujemo tudi jezikoslovno in glasoslovno poznavanje govorjenega jezika (Dobrišek 2001).

Za govorjeni slovenski jezik nekaj tovrstnega znanja že obstaja, vendar menimo, da bo to znanje dovolj dobro šele, ko bodo izvedene obsežnejše statistične analize slovenskega govorjenega jezika. Pod statistično analizo razumemo načrtno zbiranje, urejanje, predstavljanje in tolmačenje zbranih podatkov. Inženirsko delo na tem

področju se mora nanašati predvsem na pridobivanje, urejanje in predstavljanje podatkov ter manj na njihovo tolmačenje. Slednje se praviloma prepušča drugim raziskovalnim vejam, predvsem jezikoslovju, najbolj glasoslovju.

5.2 Pridobivanje govornih posnetkov

Danes sta v veljavi dva načina pridobivanja govornih posnetkov. V prvem primeru govorci izgovorijo v snemalno napravo vnaprej pripravljeno besedilo. V drugem primeru pa snemamo spontano govorjeno besedilo, ki je lahko bodisi monolog ali pogovor. Izbira besedila je ključnega pomena in je odvisna od namena zbirke.

Tudi za pridobivanje govornih posnetkov moramo izvesti načrtovanje in pripravo snemalnega okolja. Kakšne snemalne naprave in snemalno okolje izberemo, je odvisno od namena govorne zbirke. Navadno so to kar laboratorijska okolja, pri katerih pazimo na nepotrebno zvočno »onesnaževanje« (Dobrišek 2001). Primernost snemalnega okolja določa namen zbirke, vendar pogosto vseh zahtev zaradi pomanjkanja sredstev ne moremo povsem izpolniti. V splošnem pa ni potrebe, da bi govorne posnetke morali pridobivati v posebnem studijskem okolju, zato to počenjamo le v primerih, ko zbirka predstavlja dolgoročno dediščino.

Med snemanjem govora je priporočljivo preverjati, ali govorec ustrezno izgovarja predloženo besedilo. To lahko storimo s prisotnostjo druge osebe med snemanjem ali pa poskušamo v snemalni postopek vgraditi določeno samodejno preverjanje.

Govor snemamo preko mikrofona v analogni ali digitalni obliki na različne snemalne naprave. Danes to pogosto izvedemo kar na računalnikih, ki imajo vgrajeno razširitev za zajemanje zvočnih posnetkov. Posnetke govora shranjujemo v digitalni obliki na trajne računalniške pomnilniške medije. Za pridobivanje govornih posnetkov uporabljamo posebne programske uporabniške vmesnike, ki besedilo, ki ga mora izgovoriti govorec, izpišejo na zaslon računalnika ter preverjajo skladnost posnetega govora s predloženim besedilom (Dobrišek 2001).

Način snemanja govorne zbirke je odvisen od njenega namena. Če želimo proučevati lastnosti spontanega pogovora oz. razliko med govorjenimi in pisanimi besedili, snemanje opravimo v drugačnih razmerah kot denimo snemanje govorne zbirke za difonski sintetizator govora. V slednjem primeru je namreč priporočljivo, da govorec besedilo, ki vsebuje vsa želena zaporedja alofonov, prebere v celoti naenkrat (Mihelič 2002). Govorec vse besedilo izgovori na podoben način, s konstantno intonacijo. Snemanje besedila po kosih v daljšem časovnem obdobju ni priporočljivo, saj se govorcu lahko glas zaradi različnih zunanjih (vreme, drugačne nastavitve pri snemanju, spremenjen spekter in intenziteta motenj iz okolice) ali notranjih (razpoloženje, bolezen) vzrokov spremeni.

Po drugi strani pa želimo pri snemanju pogovorov oz. prostega govora zajeti čimveč prvin, po katerih se tak govor razlikuje od branega besedila, kot so značilni prekrivajoči se govor, ponavljanje, premori, zapolnjevalci vrzeli, samokorekture in napačni starti (Kranjc 1998; Stabej 2000).

5.3 Segmentacija in označevanje govorne zbirke

Posneti govorni signal predstavlja le en del govorne zbirke. Ta je brez ustreznih oznak govornih odsekov večinoma neuporabna za nadaljnje raziskave. Sledi dolgotrajni postopek segmentacije in označevanja govornega signala. Govorni signal je v postopku segmentacije potrebno razmejiti oz. segmentirati na posamezne segmente ali govorne odseke in jim v postopku označevanja ali anotacije pripisati oznake na različnih anotacijskih nivojih: grafemskem, fonetičnem, prozodijskem. Vrste segmentacije oz. oznak, ki jih govorna zbirka vsebuje, so odvisne od namena uporabe zbirke.

Za raziskave na področju govornih tehnologij moramo zbirko navadno opremiti vsaj z oznakami na grafemskem in fonetičnem nivoju. Ker je ročna segmentacija govora na fonetičnem nivoju naporna in dolgotrajna, se pri tem poslužujemo vsaj delno avtomatiziranih postopkov, ki so bolj učinkoviti, če vnaprej poznamo grafemski prepis govorjenega gradiva.

Samodejno grobo označevanje govorne zbirke

S postopkom siljenega prileganja posnetkov govora z grafi modelov glasov, ki so določeni iz fonetičnih prepisov izgovorjenih različic besed, si lahko močno olajšamo dolgotrajno in zamudno ročno segmentacijo in označevanje glasov (Dobrišek 2001). Postopek temelji na prikritih Markovovih modelih. Rezultat samodejnega siljenega prileganja so, med drugim, tudi podatki o časovnih odsekih, ki pripadajo posameznim glasovom.

Postopek siljenega prileganja posnetkov govora z grafi modelov glasov potrebuje za svoje delovanje natančno zaporedje fonemov v govornem signalu, ki ga obdeluje. Zato je potrebno grafemski prepis besedila sprva pretvoriti v fonetični prepis, ročno ali pa z uporabo samodejnega postopka za grafemsko fonetično pretvorbo, ki se uporablja tudi pri samodejni sintezi govora (Gros 2000).

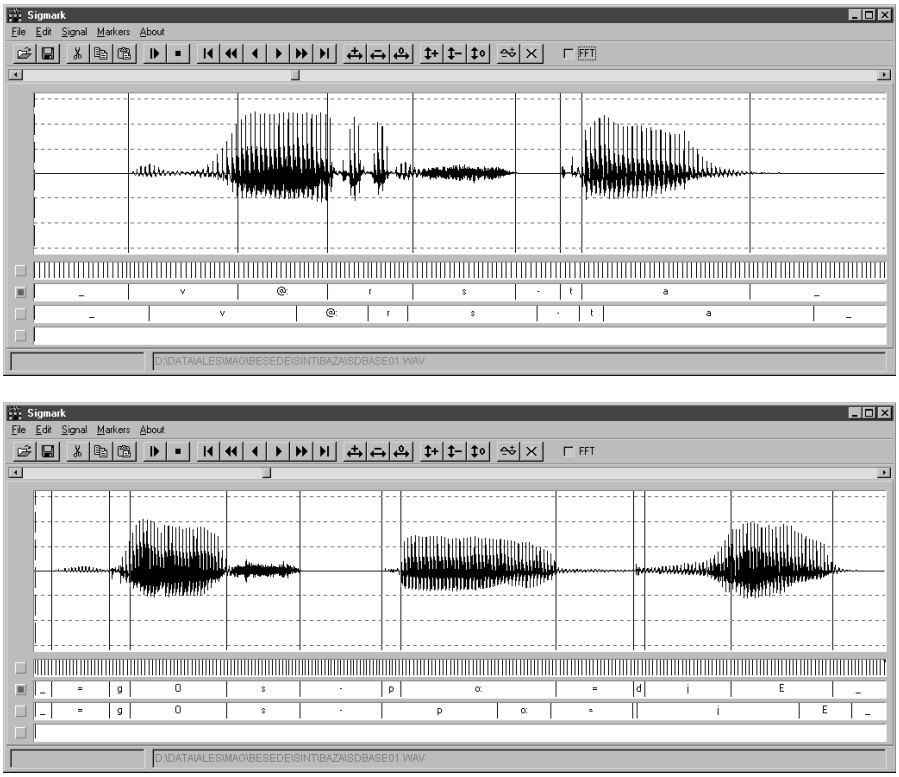
Pogosto sprva s postopkom siljenega prileganja z oznakami opremimo le manjši del govorne zbirke. Rezultat postopka nato ročno preverimo in popravimo vse napačno postavljene oznake mej med posameznimi fonemi (Mihelič 2002).

Zelo dobra lastnost omenjenega postopka za avtomatsko označevanje posnetkov je ta, da se je sposoben učiti. Vse predhodne ročne korekcije se upoštevajo pri naslednjem avtomatskem označevanju. Tako se postopek za samodejno segmentacijo in označevanje postopoma priuči načinu govorjenja govorca ter generira vse manj napak.

Fino ročno označevanje govorne zbirke

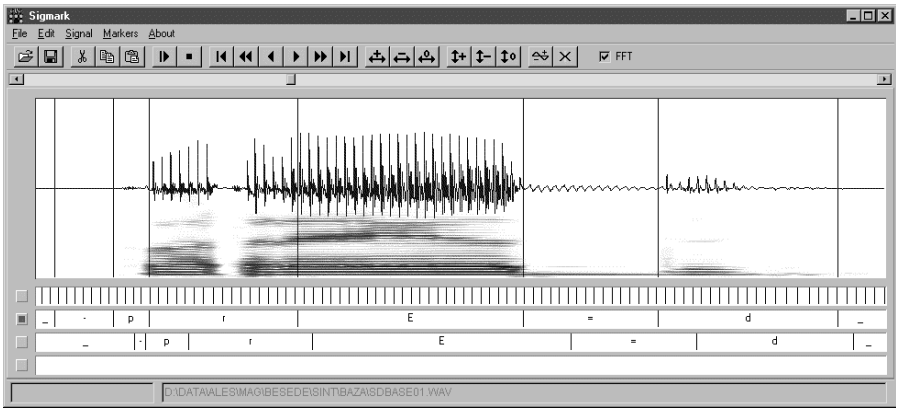
Za ročno pregledovanje in označevanje govorne zbirke ter popravljanje oznak govornih segmentov se uporabljajo raznovrstna programska orodja, namenjena delu z govornimi signali.

Kot primer takega programskega orodja predstavljamo orodje Sigmark, ki omogoča prikaz in obdelavo posnetih govornih signalov in izbranih akustičnih značilk ter poslušanje poljubnih odsekov signala.



Slika 1: *Orodje Sigmark omogoča opremljanje govornega signala z oznakami o mejah med govornimi segmenti in s simboli za označitve govornih segmentov. Prva skupina oznak predstavlja potek osnovne frekvence signala, druga skupina oznak so ročno popravljeni položaji mej med glasovi, tretja skupina pa prikazuje samodejno določene položaje mej med glasovi.*

Iz posnetkov govornih signalov je možno z metodami obdelave signalov pridobiti standardne akustične značilke govora, kot so potek osnovne frekvence, kratkočasovna glasnost govora ali prvih nekaj formantnih frekvenc govornega signala. Orodje Sigmark omogoča sočasni prikaz časovne in kratkočasovne frekvenčne karakteristike signala, kar močno olajša preverjanje in popravljanje oznak ter mej med posameznimi govornimi segmenti. Največja prednost orodja Sigmark je v tem, da omogoča samodejno in konsistentno postavljanje prozodijskih oznak za osnovno periodo.



Slika 2: *Sočasen prikaz časovnega poteka signala in kratkočasovne frekvenčne karakteristike.*

6 Pregled govornih zbirk za slovenski jezik

V tem razdelku podajamo pregled računalniško berljivih in označenih govornih zbirk za slovenski jezik. Vse omenjene govorne zbirke, razen posebnih namenskih govornih zbirk, večinoma upoštevajo osnovne zahteve po fonetični uravnoteženosti besedil in narečni pokritosti slovenskih narečnih skupin z izbranimi govorci. Bolj obsežen opis posamezne govorne zbirke se nahaja v referencah, na katere se sklicujemo ob navedbi posamezne govorne zbirke.

Sprva navajamo obsežnejše govorne zbirke, zbrane pod okriljem posamezne razvojno raziskovalne ustanove, sledi pregled namenskih govornih zbirk.

Na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru so posneli in označili vrsto govornih zbirk, ki so namenjene predvsem razvoju prepoznavanja in sinteze govora: govorna zbirka SNABI (Kačič 2002), govorna zbirka SpeechDat II (Kačič 2002), govorna zbirka PoliDat (Kačič 2002) in zbirka posnetkov emocionalnega govora (Hozjan 2002).

Na Fakulteti za elektrotehniko Univerze v Ljubljani so ustvarili naslednje govorne zbirke: govorna zbirka Mobiluz (Dobrišek 1998, Gros 2000, Mihelič 2003), govorna zbirka K211D (Dobrišek 2001), govorni zbirki radijskih in televizijskih vremenskih napovedi VNTV in VNRAD (Žibert 2000) ter štiri specializirane govorne zbirke: difonska govorna zbirka (Gros 2000), govorna zbirka VINDAT (Škrlj 2001), del zbirke Gopolis, posnet z različnimi hitrostmi govorjenja (Gros 2000), ter govorna zbirka za istovetenje govorcev (Kranjc 2001).

Na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru so zbrali govorno zbirke Števke (Rozman 2000).

Za potrebe sinteze in prepoznavanja govora so bile ob že omenjenih posnete še naslednje govorne zbirke: difonska govorna zbirka Inštituta Jožef Stefan (Šef 1998),

difonska in polifonska govorna zbirka podjetja Masterpoint (Mihelič 2002) in učna baza izgovarjav Hermes Softlab (Šket 2002).

Manjše, specializirane zbirke govorjenih besedil, namenjene predvsem jezikoslovnim raziskavam, so opisane v delih (Ozbič 1998; Modic 2002; Tivadar 2003 in Zemljak 2002).

Izmed vseh naštetih govornih zbirk sta prosto dostopni v raziskovalne namene edinole zbirki MobiLuz in K211D. Govorna zbirka SpeechDat II je dosegljiva preko distribucijske agencije ELDA <www.elda.fr>.

7 Sklepne misli

Pri raziskovalnem delu na področju tehnologij govorjenega jezika se neizbežno srečamo z zahtevo po dostopu do obsežnejših govornih zbirk. Pomanjkanje načrtno zbranih in dostopnih govornih zbirk predstavlja ključno oviro za hitrejši razvoj tehnologij govorjenega slovenskega jezika. V primerjavi z večjimi svetovnimi jeziki je za slovenski govorjeni jezik govornih zbirk razmeroma malo, pa še te so pogosto nedostopne preostalim razvojnim in raziskovalnim skupinam.

Pridobivanje govornih zbirk zahteva precej ljudi in sredstev. Zato avtorji članka pozivamo razvojno raziskovalne skupine, ki v slovenskem prostoru delujejo na področju govornih tehnologij, da se pri zbiranju govornih zbirk medsebojno povezujejo in da omogočijo dostop do lastnih govornih zbirk tudi svojim kolegom iz drugih skupin, bodisi pod razumnimi finančnimi pogoji ali pa celo brezplačno v primerih, ko se pridobljeno znanje uporablja izključno za nekomercialne raziskovalne namene.

Literatura

Dobrišek, S., Gros, J., Ipšič, I., Pepelnjak, K. Mihelič, F. Pavešić, N., 1998: Gopolis: slovenska podatkovna zbirka govorjenih poizvedovanj. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik.* Ljubljana: Institut Jožef Stefan. 105–108.

Dobrišek, S., 2001: Analiza in razpoznavanje glasov v govornem signalu. *Doktorska disertacija*. Univerza v Ljubljani: Fakulteta za elektrotehniko.

Erjavec, T., 1998: Standardizacija zapisa jezikovnih podatkov. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik.* Institut Jožef Stefan. 119–123.

Fourcin, A., Harland, G., Barry, W., Hazan, 1989: *SPEECH INPUT AND OUTPUT ASSESSMENT Multilingual Methods and Standards.* J. Wiley & Sons.

Gibbon, D., Moore, R., Winski R. (ur.), 1997: *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gros, J., 2000: *Samodejno tvorjenje govora iz besedil: postopek za izdelavo sintetizatorja slovenskega govora*. Zbirka Linguistica et philologica. Ljubljana: Založba ZRC.

Hozjan, V., Kačič, Z., Ambruš Čeh, D., 2000: Analiza prozodijskih značilk emocionalnega govora. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije*. Institut Jožef Stefan. 31–34.

Hozjan, V., Kačič, Z., 2002: Objective analysis of emotional speech for English and Slovenian Interface emotional speech databases. *Zbornik LREC–2002*. Las Palmas. 2019–2022.

Kačič, Z. in Horvat, B., 1998: Izgradnja infrastrukture potrebne za razvoj govorne tehnologije za slovenski jezik. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik*. Institut Jožef Stefan. 100–104.

Kačič, Z., Horvat, B., Zoegling Markuš, A., 2000: Issues in design and collection of large telephone speech corpus for Slovenian language. *Zbornik LREC–2002.* Atene. 943–946.

Kačič, Z., 2002: Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije*. Institut Jožef Stefan. 111–115.

Kranjc, R., 2001: *Domofonski sistem s samodejnim istovetenjem oseb na podlagi izgovorjenega gesla*. Diplomsko delo. Univerza v Ljubljani: Fakulteta za elektotehniko.

Kranjc, S., 1998: Govorjena besedila in korpus slovenskega jezika. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik.* Institut Jožef Stefan. 109–112.

Marascuilo, M. in Serlin, R., 2002: *Statistical methods for the social, and behavioral sciences.* Freeman and Company. New York.

Mihelič, F., Gros, J., Noeth, E., Dobrišek, S., Žibert, J., 2000: Recognition of Selected Prosodic Events in Slovenian Speech. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije*. Institut Jožef Stefan. 45–49.

Mihelič, A, Gros, J., Pavešić, N., Žganec, M., 2002: Pridobivanje govorne zbirke za korpusni sintetizator govora Phonectic. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije*. Institut Jožef Stefan. 14–19.

Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., Pavešić, N., 2003: Spoken Language Resources at LUKS of the University of Ljubljana. *International Journal of Speech Technology* 6/3. 221–232.

Modic, R. in Petek, B., 2002: A Contrastive Acoustic Phonetic Analysis of Slovenian and English Diphthongs. *Zbornik LREC–2002.* Las Palmas. 293–296.

Ozbič, M., 1998: Akustična spektralna FFT analiza samoglasniškega sistema slovenskega jezika. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik*. Institut Jožef Stefan. 55–59.

Rojc, M. in Kačič, Z., 2000: Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system. *Zbornik LREC–2002.* Atene. 321–325.

Rozman, R., Kodek, D., 2000: Govorna baza »Števke« in raziskave robustnosti sistemov za razpoznavanje govora. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije.* Institut Jožef Stefan. 75–78.

Srebot-Rejec, T., 1998: *Word Accent and Vowel Duration in Standard Slovene: An Acoustic and Linguistic Investigation.* Slavistische Beiträge 226. München: Verlag Otto Sagner.

Stabej, M., Vitez, P., 2000: KGB (korpus govorjenih besedil) v slovenščini. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije.* Institut Jožef Stefan. 79–81.

Šef, T., Dobnikar, A., Gams, M., Grobelnik, M., 1998: Slovenski govor na internetu. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije za slovenski jezik.* Institut Jožef Stefan. 60–64.

Šket, G. in Imperl, B., 2002: M-vstopnica – uporaba avtomatskega razpoznavanja govora v praksi. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije.* Institut Jožef Stefan. 116–119.

Škrlj, M., Mihelič, F., Gros, J., Dobrišek, S., 2001: Speech corpora VINDAT – The Influence of the Psychophysical Condition of the Speaker on Speech Characteristics. *Zbornik ERK–2001* Portorož. 261–264.

Tivadar, H., 2003: *Govorjena podoba slovenskega knjižnega jezika – pravorečni vidik.* Magistrsko delo. Univerza v Ljubljani: Filozofska Fakulteta.

Toporišič, J., 1992: *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba.

Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P., 2002: Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija* 50/2. 159–169.

Zemljak, M., 2002: *Trajanje glasov štajerskega zabukovškega govora: instrumentalno-slušna analiza.* Dokorska disertacija. Univerza v Ljubljani: Filozofska Fakulteta.

Zögling Markuš, A. Kačič, Z. Horvat, B., 2000: Razvoj slovenske baze izgovarjav »POLIDAT«. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije.* Institut Jožef Stefan. 95–98.

Žibert, J., Mihelič, F., 2000: Govorna zbirka vremenskih napovedi. Erjavec, T. in Gros, J. (ur.): *Zbornik konference Jezikovne tehnologije.* Institut Jožef Stefan. 108–111.

Pomembnejše spletne strani

LUKS <http://www.luks.fe.uni-lj.si>
FERI <http://www.dsplab.uni-mb.si/>
FRI <http://mrl-pc.fri.uni-lj.si/>
IJS <http://nl.ijs.si>
Masterpoint <http://<www.masterpoint.si>
HSL <http://www.hermes-softlab.com/>
Alpineon <http://www.alpineon.com>
FIDA <http://www.fida.net>)
Nova beseda <http://bos.zrc-sazu.si/s_beseda.html>