Govorna zbirka vremenskih napovedi

Janez Žibert*, France Mihelič*

*Laboratorij za umetno zaznavanje
Fakulteta za elektrotehniko
Univerza v Ljubljani
Tržaška 25, 1001 Ljubljana, Slovenija
{janez.zibert, france.mihelic}@.fe.uni-lj.is

Povzetek

Predstavljena bo govorna zbirka vremenskih napovedi v slovenskem jeziku. Govorna zbirka je prvi poskus zbiranja govornih posnetkov iz televizijskih in/ali radijskih oddaj. Namen zbirke je uporaba za različne aplikacije razpoznavanja (delno) spontanega slovenskega govora. V članku je podano nekaj lastnosti govorne zbirke, predstavljena so orodja za označevanje in zbiranje posnetkov. Opisan pa je tudi razpoznavalnik govora zgrajen na podlagi zbirke, ki smo ga uporabili za avtomatsko označevanje posnetkov.

1. Uvod

Govorne tehnologije postajajo v zadnjih letih področje intenzivnega raziskovanja številnih znanstvenikov iz različnih področij. Vse bolj izrazita je namreč potreba po inteligentnih strojih, ki se bodo znali prilagajati človekovim željam in potrebam ter bo njihovo upravljanje čimbolj enostavno. To pa pomeni, da se človek ne bo učil zapletenih postopkov upravljanja takšnih strojev, ampak se bo stroj prilagajal človekovem ravnanju in razmišljanju. Prva značilnost takšnih inteligentnih strojev pa je prav gotovo način komunikacije s človekom. In ker je govor človeku najbolj naravno sredstvo komuniciranja, se v zadnjem času veliko pozornost posveča ravno strojnemu razpoznavanju in tvorjenju govora.

Skupne značilnosti postopkov za razpoznavanje in tvorjenje govora v različnih jezikih so dobro pripravljene in označene govorne baze. Govorna baza predstavlja namreč osnovo tako za izgradnjo razpoznavalnikov kot postopkov za tvorjenje govora. Dobro pripravljena in označena baza je zato prvi korak v smeri robustnega in čimbolj zanesljivega razpoznavanja govora stroja v komunikaciji s človekom. Poleg tega je ravno govorna baza tista, ki naj bi zajela čim več značilnosti posameznega jezika in je zato zelo specifična za posamezen jezik. V veliki meri tako predstavlja najbolj občutljivi in hkrati najpomembnejši člen v izgradnji razpoznavalnika govora za točno določen jezik.

Statistične metode, ki so se uveljavile pri obdelavi in analizi govora, so poudarile pomen govornih zbirk, ki so postajale vse večje in kompleksnejše. V devetdesetih letih je bilo narejeno veliko dela in predstavljenih veliko aplikacij za avtomatično zbiranje in transkripcijo televizijskih in radijskih govorjenih oddaj (Kubala, 1996). Te namreč predstavljajo skoraj neomejen vir govornih podatkov.

Govorni posnetki zbrani iz radijskih in/ali televizijskih oddaj v splošnem niso homogeni. Vsebujejo lahko različne tipe govora: bran ali spontan govor, pogovorni ali zborni jezik, različne tipe govornega signala (telefonski govor, studijski posnetki) s prepletanjem glasbe ali drugih šumov iz ozadja. Zato je potrebno vnaprejšnje načrtovanje zbirke in skrbno označevanje posnetkov.

V članku bomo opisali govorno bazo vremenskih napovedi, ki predstavlja prvi poskus govorne zbirke na podlagi televizijskih in/ali radijskih oddaj v slovenskem jeziku. Dosedanje slovenske govorne zbirke namenjene razpoznavanju (sintezi) govora so omejene zgolj na brani govor, (Dobrišek, 1998), (Kačič, 2000).

Predstavljeno bo zbiranje in označevanje posnetkov, podali pa bomo tudi nekaj rezultatov razpoznavanja posnetkov vremenskih novic.

2. Kako je nastajala govorna zbirka

Odločitev za vremenske napovedi je pogojena z namenom uporabe zbirke in sorazmerno manjšim obsegom obdelave posnetkov, predstavlja pa prvi korak v smeri snemanja večjih govornih baz televizijskih in radijskih novic (Garofolo, 1997).

Govorna zbirka se loči na dva dela: na televizijske vremenske napovedi VNTV in radijske vremenske napovedi VNRAD.

Posnetke za govorni bazi *VNTV* in *VNRAD* smo zbirali od konca oktobra 1999 do konca marca 2000. Govorna baza *VNTV* zajema posnetke televizijskih vremenskih napovedi nacionalne televizije TVSLO1, in sicer vsak dan po tri vremenske napovedi istega govorca, medtem ko baza *VNRAD* vključuje posnetke jutranjih radijskih vremenskih napovedi, ki so bili posneti prav tako vsak dan na nacionalni radijski postaji.

Dnevna snemanja televizijskih napovedi so potekala z ATI-jevo All in Wonder grafično kartico z vgrajenimi televizijskim sprejemnikom. Pogoji snemanja so bilo vedno približno enaki, tako da so posnetki enako kvalitetni. Radijske napovedi so bile snemane z radijsko kartico prav tako ob vedno istih pogojih. Posnetki so vzorčeni s frekvenco 22050 Hz in shranjeni v 16-bitnem PCM mono wav formatu (Windows WAVE header).

Posnetki vremenskih napovedi so shranjeni v celoti kot tudi razrezani po posameznih stavkih. Končna verzija zbirke bo izšla na CD-ROMih.

Slika 1: Označevanje govornih posnetkov z orodjem Transcriber (trans, 1999).

3. Označevanje posnetkov

Pri govornih posnetkih vremenskih napovedi ločimo dve obliki govora: spontan in načrtovan govor. Govor pri televizijskih vremenskih napovedih, kjer si govorci že vnaprej pripravijo vremensko poročilo in ga delno berejo, smo označili kot načrtovan. Pri radijskih vremenskih napovedih pa je govor bolj spontan in poteka v obliki dialoga (radijski voditelj sprašuje vremenoslovca po vremenu), zato smo ga označili kot spontan govor. Razlika med oblikama je predvsem v tem, da spontan govor vsebuje večje število medmetov, vdihov, izdihov, zatikanj ipd.

Transkripcija vremenskih napovedi je potekala v dveh fazah. V prvem delu smo se s Hidrometeorološkim zavodom Republike Slovenije dogovorili, da nam pošiljajo osnutke vnaprej pripravljenih vremenskih poročil, ki jih bodo povedali na televizijski oddaji. Ti osnutki so bili dobra osnova za izdelavo transkripcij po besedah televizijskih vremenskih napovedi.

Označevanje posnetkov je poleg transkripcij po besedah vsebovalo še informacijo o začetku in koncu posameznega stavka (kot pomenske enote) ter zaporedno številko stavka. Slednje informacije smo uporabili pri rezanju napovedi na stavke. Razrez je potekal po zaključenih pomenskih enotah oziroma tam, kjer je izrazit premor med posameznimi deli govornega signala. Ob tem smo se poskušali izogniti kar največjemu številu nepravilnosti (medmetov, vdihov, izdihov, ...) v govornem signalu, ki nastopajo med posameznimi deli govora. Pri transkripciji televizijskih posnetkov smo uporabili posebne oznake v oklepajih $< . >$, s katerimi smo označevali dele govora, ki ne nosijo nobene pomenske informacije, kot npr. medmete, vdihe ali izdihe, momljanje, tleske z jezikom in podobno.

Pri označevanju posnetkov smo uporabljali orodje za transkripcijo govornega signala Transcriber (trans, 1999).

Transcriber na sliki 1 je grafično orodje za obdelavo in označevanje daljših govornih posnetkov. Namenjen je predvsem označevanju televizijskih in/ali radijskih govorjenih oddaj, saj omogoča hitro in natančno označevanje delov posnetkov, kjer je npr. glasba v ozadju, kjer se prepletata dva ali več govorcev, ipd. Transkripcije posnetkov so v SGML formatu.

Prva faza označevanja posnetkov je potekala od konca oktobra 1999 do sredine januarja 2000. Tako označeni posnetki so predstavljali učno bazo pri izgradnji razpoznavalnika govora, ki je bil uporabljen za drugo fazo označevanja govornih posnetkov vremenskih napovedi.

Druga faza je potekala že polavtomatsko. Posnete napovedi smo namreč v celoti razpoznavali z zgrajenim razpoznavalnikom. Poleg transkripcij besed pa smo poskušali avtomatič no označevati tudi premore med posameznimi deli govora, da bi postavili meje med stavki. Tako nam je preostala le še kontrola označenih posnetkov.

Proces označevanja govornih posnetkov smo tako znatno pohitrili, kar pa je bil tudi eden izmed naših ciljev.

Pri označevanju radijskih napovedi za bazo *VNRAD* smo še v prvi fazi označevanja posnetkov. Zaradi manjše količine zbranega materiala in večje spontanosti govora pri teh napovedih ter dodatnih govorcev je bil razpoznavalnik zgrajen iz televizijskih posnetkov preveč nezanesljiv, da bi ga lahko uporabili. Tu imamo opravka s še večjo količino delov govora brez pomenske informacije, za katere uporabljamo enake oznake kot pri televizijskih napovedih.

V nadaljnji fazi bomo poskušali tudi radijske posnetke razpoznavati z dodatno naučenim razpoznavalnikom, da bi pohitrili označevanje.

4. Statistika govorne zbirke

Kot smo že omenili je zbirka razdeljena na bazo televizijskih *VNTV* govornih posnetkov in radijskih vremenskih napovedi *VNRAD*.

govorec	# napovedi	# stavkov	# besed	čas trajanja
01f	36 (30/6)	789 (683/106)	7609	51 min (43.5/7.5)
01m	43 (21/22)	1078 (553/525)	12008	70 min (35/35)
02m	32 (23/9)	578 (429/149)	6398	39 min (29/10)
03m	39 (18/21)	965 (479/486)	10041	59 min (28.5/30.5)
04m	28 (20/8)	472 (349/123)	5221	32 min (23.5/8.5)
Skupaj	178 (112/66)	3882 (2493/1389)	41277 (različnih 2857)	252 min (159/93)

Tabela 1: Statistika govorne zbirke VNTV. Prva vrednost v okroglem oklepaju predstavlja učni del zbirke, druga pa testni.

Pri bazi *VNTV* imamo pet govorcev: ena ženska in 4 moški, pri *VNRAD* pa je 9 moških govorcev, od tega so štirje skupni za obe bazi.

V bazi *VNTV* so zbrani vsakodnevni posnetki treh vremenskih napovedi, ki trajajo približno od ene do dveh minut. V zbirki je 178 vremenskih napovedi, kar predstavlja približno 252 minut govornega materiala. Bazo *VNRAD* sestavlja 62 vremenskih napovedi ali približno 87 minut govornih posnetkov.

Ker označevanje radijskih posnetkov še ni končano, je vsa nadaljnja statistika narejena samo za bazo *VNTV*. Korpus stavkov predstavlja 3882 (različnih) stavkov. Slovar je sestavljen iz 2857 besed.

Pri zbiranju podatkov iz televizijskih in/ali radijskih govorjenih oddaj ne moremo predvidevati omejenega stavčnega in besednega korpusa in vsak nov posnetek nam lahko tako v korpus prispeva nove besede ali stavke. Ravno zato se je potrebno omejiti na eno področje ali temo, ki jo pokrivajo govorni posnetki. Zbiranje posnetkov smo končali, ko smo ugotovili, da vremenske napovedi ne prispevajo več kot 5% novih besed v slovar. Poudariti pa velja, da smo zbirali posnetke vremenskih napovedi v zimskih mesecih in slovar tako zagotovo nezadostno pokriva vremenske napovedi za toplejši del leta, kar posledično vpliva na rezultate razpoznavanja.

Zbirko *VNTV* smo nadalje razdelili na učni in testni del. Testni del je sestavljen iz 1389 stavkov (93 minut govornih posnetkov), kar predstavlja 36 % celotne zbirke.

Osnovni podatki zbirke *VNTV* po posameznem govorcu so predstavljeni v tabeli 1.

Govorna zbirka je namenjena razmeroma ozkemu področju uporabe - razpoznavanju in avtomatični transkripciji vremenskih napovedi. Sodi pa med govorne baze s srednje velikimi slovarji, razmeroma velikim stavčnim korpusom in z manjšim številom govorcev.

5. Rezultati razpoznavanja

Prvi del govorne zbirke *VNTV* smo uporabili za izgradnjo razpoznavalnika, ki smo ga uporabljali za avtomatsko označevanje posnetkov v drugi fazi. Po končanem zbiranju posnetkov pa smo razpoznavalnik doučili s posnetki iz druge faze.

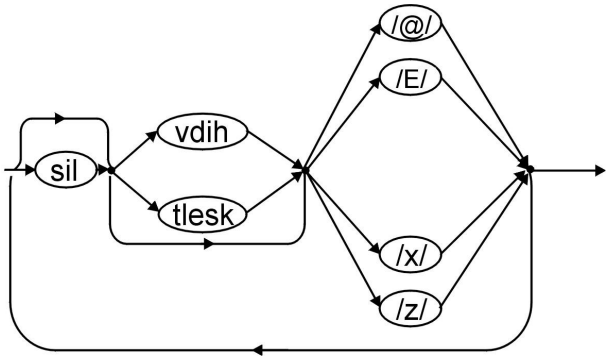
V tem poglavju bo predstavljen razpoznavalnik govora namenjen razpoznavanju posnetkov vremenskih napovedi, ki smo ga že uspešno uporabili za izgradnjo aplikacije za avtomatično podnaslavljanje televizijskih vremenskih napovedi (Žibert, 2000).

Razpoznavalnik, ki temelji na prikritem Markovovem modelu (PMM) (Rabiner, 1989), je zgrajen na standarden način.

Vektor značilk uporabljen za učenje modela in razpoznavanje je sestavljen iz logaritma energije in dvanajstih koeficientov melodičnega kepstruma s prvimi in drugimi odvodi. Uporabljen je tudi postopek ničenja srednjih vrednosti kepstruma.

Osnova za izgradnjo razpoznavalnika je zvezni prikrit Markovov model (PMM) z vezanimi stanji. Uporabili smo 5 stanj na model z dovoljenimi preskoki. Z njimi smo modelirali posplošene kontekstno odvisne modele trifonov. Z združevanjem trifonov s podobnim levim in desnim kontekstom smo tvorili skupne vezane modele. V vsakem stanju smo modelirali kombinacijo treh zveznih funkcij gostot verjetnosti z diagonalnimi kovariančnimi matrikami.

Dodatno smo zgradili še posebne modele za medmete, vdihe in izdihe govorca ter tleske z jezikom. Poseben model smo zgradili tudi za določevanje mej med stavki, kjer smo dodatno razširili in spremenili model premora med besedami. Zgrajen pa je bil tudi model za razpoznavanje besed, ki niso iz slovarja. Ta model, prikazan na sliki 2, je dejansko PMM model sestavljen iz modelov posameznih fonemov.



Slika 2: PMM model za razpoznavanje besed, ki jih ni v slovarju. Simboli v vozliščih modela predstavljajo manjše PMM modele za posamezne govorne enote (vdih , izdih, fonemi,...).

Dodatni modeli so potrebni predvsem zaradi dejstva, da smo hoteli zgraditi robusten razpoznavalnik, ki bo uporaben za različne govorne aplikacije.

Parametre tako zgrajenega PMM-ja smo določili iz učnega dela govorne zbirke *VNTV*.

Pri gradnji razpoznavalnika smo nadalje uporabili bigramski jezikovni model, določen iz stavčnega korpusa naše zbirke, ki vsebuje 41k besed. Perpleksnost tako

naučenega jezikovnega modela je 22.7.

Za gradnjo in učenje akustičnega in jezikovnega modela smo uporabljali orodje HTK (Young, 1997).

Pri testiranju razpoznavanja govora smo uporabili standarden pristop. Mera za odstotek pravilnosti razpoznanih besed v testni bazi je podana z naslednjo zvezo:

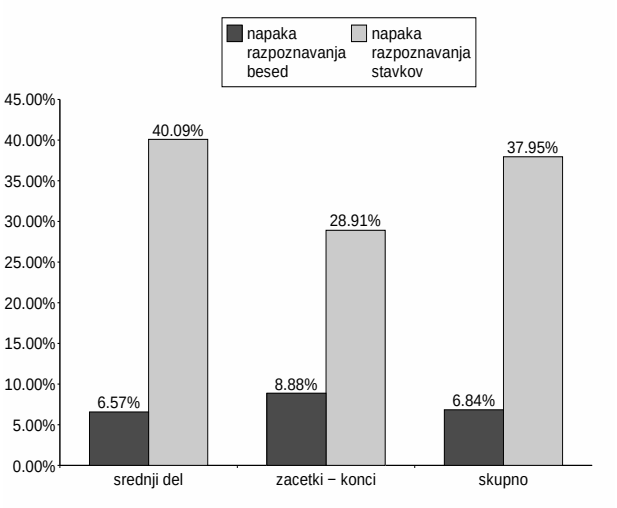
$$WA = (1 - \frac{D + S + I}{N}) * 100\%.$$

Rezultati so posledica ujemanja med posameznimi nizi (besedami, stavki), ki jih razpoznavamo, in razpoznanimi nizi (besedami, stavki), v odvisnosti od števila vrivanj (I), brisanj (D) in zamenjav (S) posameznih besed. Pri natančnosti razpoznavanja govora upoštevamo razmerje med številom pravilno razpoznanih besed ($N - (D + S + I)$) in številom vseh besed (N), ki nastopajo v posameznih stavkih testne množice. Pri pravilnosti razpoznavanja pa se izračuna razmerje med številom pravilno razpoznanih znakov brez števila vrivanj in številom vseh znakov.

V tabeli 2 so prikazani rezultati razpoznavanja za posameznega govorca.

govorec	pravilnost (%)	natančnost (%)	pravilnost stavkov (%)
01f	96.32	93.51	75.24
01m	93.23	89.41	60.15
02m	93.53	89.67	57.52
03m	92.92	90.21	57.07
04m	93.81	90.10	60.23

Tabela 2: Rezultati razpoznavanja testnega dela zbirke *VNTV*.



Slika 3: Napaka razpoznavanja po besedah in po stavkih testnega dela zbirke *VNTV*.

Napaka razpoznavanja po besedah in po stavkih testnega dela zbirke *VNTV* je prikazana na sliki 3. Tu lahko opazimo razliko v natančnosti razpoznavanja med začetnimi in srednjimi deli vremenskih napovedi. Razpoznavanje govornih posnetkov srednjih delov napovedi je boljše, saj je jezik tu strokovni, znanstveni, zato je tudi jezikovni model bolj učinkovit. Drugi razlog pa je v tem, da se v teh delih napovedi uporablja več strokovnih izrazov, ki se večkrat ponavljajo in je precej manj novih besed, zato tudi slovar zadostno pokriva te dele napovedi. Posledično pa to pomeni boljše razpoznavanje v primerjavi z začetki in konci napovedi, kjer imamo v povprečju več novih besed na napoved.

6. Zaključek

V članku je bila opisana nova govorna zbirka vremenskih napovedi, ki predstavlja prvi poskus zbiranja posnetkov iz TV in/ali radijskih oddaj v slovenskem jeziku.

Namen zbiranja takšnih govornih posnetkov je predvsem v tem, da bi pridobili znanje in izkušnje s pridobivanjem večjih govornih zbirk iz TV/radijskih oddaj. Pripravili smo vsa potrebna orodja in okolje za zbiranje takšnih podatkov. Drugi cilj zbirke je, da bi poskušali bolje razpoznavati delno spontani tekoči slovenski govor.

Odločitev za pridobivanje govornih posnetkov vremenskih napovedi pa je bila pogojena tudi z namenom uporabe zbirke. Zbirko smo že uspešno uporabili za avtomatično transkripcijo napovedi in poskusno za avtomatično sinhronizirano podnaslavljanje TV vremenskih napovedi.

Predstavljena govorna zbirka je pomembna za nadalnje raziskovalno delo na področju razpoznavanja in prozodično analizo slovenskega govora, dodatno označeno pa jo lahko uporabimo tudi za sintezo govora v sistemih za avtomatsko podajannje vremenskih informacij.

7. Zahvala

Za sprotno pošiljanje osnutkov televizijskih vremenskih poročil, ki so nam služili kot osnova za transkripcije posnetkov v prvi fazi označevanja govorne zbirke, se zahvaljujemo Hidrometeorološkemu zavodu Republike Slovenije.

8. Literatura

S. Dobrišek, J. Gros, F. Mihelič, N. Pavešić. 1998. Recording and labeling of the GOPOLIS Slovenian speech database. *Proc. 1st Int.Conf. on Language Resources & Evaluation*, 2:1089–1096.

J. Garofolo, J. G. Fiscus, W. M. Fisher. 1997. Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora. *Proceedings of DARPA Speech Recognition Workshop*.

Z. Kačič, B. Horvat, A. Zögling. 2000. Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. *Proceedings of LREC 2000, 2nd International Conference on Language Resources & EvaluationProc*, 943–946.

F. Kubala. Feb. 1996. Toward automatic recognition of broadcast news. *Proc. DARPA Speech Recognition Workshop*.

L. R. Rabiner. Feb. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, vol. 77, 2:257–286.

S. Young, J. Odell, J. Ollason, V. Vatchev, P. Woodland. 1997. The HTK Book. Cambridge University, Entropic Cambridge Reasearch Laboratory Ltd.

Transcriber: http://www.etca.fr/CTA/gip/Projets/Transcriber/.

J. Žibert, F. Mihelič. 2000. Avtomatično podnaslavljanje vremenskih napovedi. *Zbornik devete Elektrotehniške in računalniške konference ERK 2000*.