Govorni in jezikovni viri slovenščine za samodejno razpoznavanje tekočega govora

Gregor Donaj, Andrej Žgank, Mirjam Sepesy Maučec
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Smetanova ul. 17, 2000 Maribor
gregor.donaj@um.si, andrej.zgank@uni-mb.si, mirjam.sepesy@uni-mb.si

Izvleček

Govor je za ljudi najbolj naravno komunikacijsko sredstvo. Govorno komunikacijo s strojem omogočajo sistemi za samodejno razpoznavanje govora. Različne aplikacije razpoznavanja govora so za stroj različno zahtevne. Med najzahtevnejše štejemo samodejno razpoznavanje tekočega govora. Aplikacije razpoznavanja govora temeljijo na statistični obdelavi govornega signala ter gradnji akustičnih in jezikovnih modelov. Za izdelavo teh modelov je pomembna uporaba kakovostnih govornih in jezikovnih virov. V prispevku opisujemo govorne in jezikovne vire za slovenščino, ki se uporabljajo za samodejno razpoznavanje govora. Predstavimo tudi modularno zgradbo razpoznavalnika. V eksperimentalnem sistemu analiziramo vpliv uporabe modelov v razpoznavalniku tekočega govora v domeni dnevnoinformativnih oddaj.
Ključne besede: govorni viri, jezikovni viri, akustični modeli, jezikovni modeli, samodejno razpoznavanje govora.

Abstract

Slovene Speech and Language Resources for Automatic Speech Recognition

Speech is the most natural way of communicating. Speech communication with machines is made possible with systems for automatic speech recognition. Different applications of speech recognition are differently challenging. Among the most challenging is continuous speech recognition. Speech recognition systems are based on statistical speech signal processing and the building of acoustical and language models. Quality speech and language resources are needed to build these models. This paper gives an overview of speech and language resources for Slovene, which are used in automatic speech recognition. A modular structure of a speech recognizer is also presented. In an experimental system the impact of using different models on the accuracy in a Broadcast News speech recognition system is analyzed.
Key words: speech resources, language resources, acoustical models, language models, automatic speech recognition.

1 UVOD

Govor kot človekovo najbolj naravno komunikacijsko sredstvo pomeni za stroj zelo kompleksno nalogo. Razpoznavanje tekočega govora in razpoznavanje spontanega govora sta za raziskovalce polna izzivov. Posebnosti posameznih jezikov razpoznavanje govora še dodatno zapletejo. Tudi slovenščina kot visoko pregibni jezik spada v skupino bolj zahtevnih jezikov za razpoznavanje.

Poznamo različne pristope samodejnega razpoznavanja govora (angl. Automatic Speech Recognition, ASR). Med preprostejše štejemo razpoznavanje izoliranih besed z majhnim slovarjem, med zahtevnejše pa razpoznavanje tekočega govora z velikim slovarjem (Sepesy Maučec, Rotovnik, Kačič & Brest, 2009). Za obe aplikaciji je pomembno, da imamo izdelane dobre modele govora. V primeru razpoznavanja izoliranih besed so predvsem pomembni akustični modeli, ki modelirajo akustične značilnosti govora. Ti modeli služijo prepoznavanju fonemov in besed. Razpoznavanje tekočega govora pa pomeni še večjo zahtevnost za akustično modeliranje, saj je treba upoštevati tudi prehode med besedami, ker so v tekočem govoru zabrisane meje med besedami. Dodatno so pri razpoznavanju tekočega govora velikega pomena statistični jezikovni modeli. Z njimi modeliramo verjetnosti zaporedij besed v jeziku. Pri izdelavi jezikovnih modelov se pogosto poslužujemo pisnih virov jezika. Posledično so jezikovni modeli bolj primerni za razpoznavanje branega govora, manj pa za razpoznavanju spontanega govora (Žgank & Sepesy Maučec, 2010).

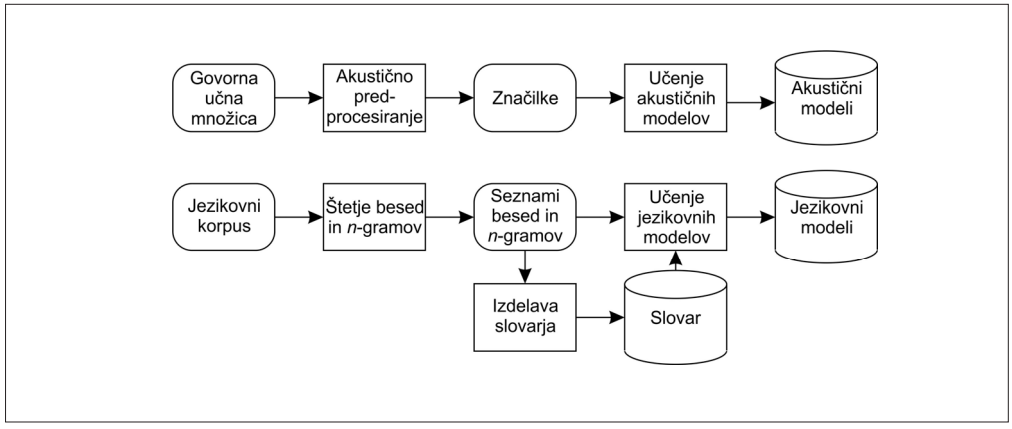
Tako za izdelavo akustičnih kot jezikovnih modelov so pomembni kakovostni in dovolj obsežni govorni oz. pisni viri jezika. V članku bomo predstavili nekatere takšne vire, ki so na voljo za slovenski jezik.

Njihovo uporabnost bomo predstavili na primeru razpoznavalnika tekočega govora UMB Broadcast News, ki je bil razvit na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru.

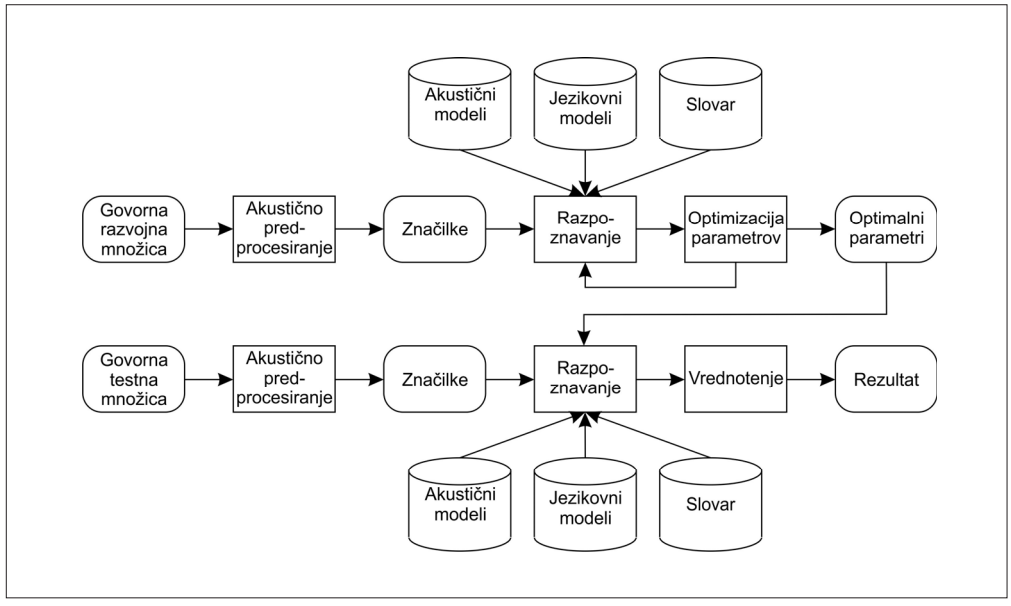
V drugem razdelku bomo predstavili osnovno zgradbo in module sistema za ASR. V tretjem razdelku bomo opisali posebnosti slovenščine, zaradi katerih je ta za razpoznavanje govora večji izziv. Sledi opis osnovnih govornih in jezikovnih virov za slovenščino, ki so uporabni za gradnjo sistemov ASR. V četrtem razdelku je opisan eksperimentalni sistem, v petem razdelku pa rezultati eksperimentov. V šestem razdelku sledi sklep.

2 SAMODEJNO RAZPOZNAVANJE GOVORA

Delovanje sistemov za samodejno razpoznavanje govora delimo na dve fazi. Prva faza je učenje jezikovnih in akustičnih modelov. Blokovna shema učenja modelov je prikazana na sliki 1. Končni rezultat te faze so akustični in jezikovni model ter slovar besed.



Slika 1: **Postopek učenja akustičnih in jezikovnih modelov**



Slika 2: **Delovanje razpoznavanja govora**

Druga faza je razpoznavanje. Njena blokovna shema je prikazana na sliki 2. Sistem za razpoznavanje govora na vhodu sprejme zvočni signal, na izhodu pa posreduje razpoznano zaporedje besed. Sistem ima modularno zgradbo, module pa lahko razdelimo v dve skupini: na module za predprocesiranje

govora in module za razpoznavanje govora. Vhodni zvočni signal najprej obdela modul za akustično segmentacijo, ki zvočni signal razdeli na akustično homogene dele. Modul za akustično analizo izlušči informacijo v govoru in jo predstavi z vektorjem akustičnih značilk. Postopek izločanja značilk mora biti popolnoma enak kot pri učenju akustičnih modelov. Niz vektorjev značilk je vhodni podatek iskalnega algoritma, ki poišče najbolj verjetno zaporedje izgovorjenih besed. Pri tem uporablja informacijo iz akustičnih in jezikovnih modelov. Akustični modeli opisujejo akustične lastnosti govora na ravni fonemov, jezikovni modeli pa jezikovne lastnosti govora na ravni besed. Oboji, tako akustični kot jezikovni modeli, temeljijo na statističnem procesiranju govora oz. jezika. Razpoznavanje na razvojni množici poteka z namenom iskanja optimalnih parametrov razpoznavanja – uteži akustičnih in jezikovnih modelov. Končni rezultat uspešnosti razpoznavanja dobimo na testni množici, pri čemer uporabimo optimizirane vrednosti parametrov.

2.1 Akustični modeli

Akustični modeli so ključni gradnik samodejnega razpoznavalnika govora s stališča procesiranja govornega signala. Njihova naloga je modelirati akustično-fonetične lastnosti govora, pri tem pa v primeru razpoznavanja govora neodvisnega govorca uspešno zmanjšati razlike med posameznimi govorci. Osnovna enota akustičnih modelov je običajno fonem, ki ga zaradi modeliranja učinka koartikulacije modeliramo v širšem kontekstu predhodnega in naslednjega fonema. Takšen akustični model poimenujemo trifon. Na trifon lahko gledamo kot na posplošitev pojma alofon. Alofoni so različne možne izgovorjave nekega fonema glede na njegov kontekst. Za vsak fonem imamo običajno le majhno množico alofonov. Definicija trifona pa zajema vse možne kombinacije treh zaporednih fonemov (za *N* fonemov pomeni to *N*3 trifonov). Medtem ko definicija alofona izhaja iz fonologije, pa trifone uvajamo v obdelavi govora zaradi zveznih sprememb vokalnega trakta, ki nastopijo pri prehodu iz izgovarjave enega fonema na naslednjega in se odražajo v akustičnem signalu govora ob tem prehodu. Primer fonetične in grafemske oblike vnosa besede »avtomatskega« v slovarju razpoznavalnika govora je prikazan v tabeli 1.

Tabela 1: **Primer fonetične in grafemske oblike vnosa v slovar razpoznavalnika govora**

Beseda	Kategorija transkripcije	Transkripcija
avtomatskega	MRPA fonemi	a U t O m “a: ts k E g a
avtomatskega	Grafemi	a v t o m a t s k e g a

Za akustično modeliranje pri ASR se uporabljajo različni pristopi (Aubert, 2002), najpogostejši so prikriti modeli Markova (angl. Hidden Markov Model, HMM), uteženi končni pretvorniki (angl. Weighted Finite State Transducer, WFST) in nevronske mreže (angl. Artificial Neural Network, ANN). V predstavljenem eksperimentu smo uporabljali tristanjske levo-desne prikrite modele Markova z zveznimi Gaussovimi porazdelitvenimi funkcijami verjetnosti. Za slovenski jezik je pretvorba med grafemi in fonemi netrivialen proces, ki lahko k rezultatom razpoznavanja govora vnese dodatno napako.

2.2 Jezikovni modeli in slovarji

Pri razpoznavanju govora so meje med besedami zabrisane, saj v tekočem govoru med besedami ni premorov. Za določanje zaporedja besed so najprej uporabljali deterministične besedne mreže, ki so jih nasledili jezikovni modeli, temelječi na pravilih slovnice jezika. Sestavljanje slovničnih pravil, ki bi pokrila jezik kot celoto, je zelo zahtevna naloga, ki zahteva poglobljeno znanje o jeziku. Po drugi strani pa imamo v spontano govorjenem jeziku veliko slovnično nepravilnih zaporedij. Ideja jezikovnega modela je določiti verjetnost poljubnemu zaporedju besed. Jezikovni model lahko obravnavamo tudi kot model, ki v procesu razpoznavanja napoveduje najbolj verjetno naslednjo besedo. Za jezikovni model velja tudi to, da verjetnost zaporedja besed ni nikoli enaka nič, kar je še posebno dobrodošlo pri razpoznavanju spontanega govora. V praksi so se najbolj uveljavili statistični n-gramski jezikovni modeli, ki verjetnost poljubnega zaporedja besed izračunajo s sestavljanjem verjetnosti n-gramov. V jezikovnih modelih označuje n-gram zaporedje *n* besed, *n* pa določa red n-grama. Najpogostejši so bigrami (2-grami) in trigrami (3-grami), zasledimo pa tudi uporabo jezikovnih modelov do reda 5 (tj. 5-gramov). Smiselnost uporabe jezikovnih modelov višjih redov je povezana z velikostjo učnega korpusa, tj. besedila, v katerem štejemo modelirane n-grame. Da je verjetnost poljubnega zaporedja besed vedno večja od 0,

zagotavljajo metode glajenja verjetnosti (Chen & Goodman, 1999), ko določeno, resda majhno, verjetnost pripišejo tudi n-gramom, ki se nikoli ne pojavijo v učnem korpusu. Preliminarne raziskave so pokazale, da je za modeliranje slovenskega jezika najučinkovitejše glajenje, ki temelji na Good-Turingovem glajenju (Good, 1953) in sestopanju po Katzu (1987).

Jezikovni modeli opisujejo verjetnostne lastnosti n-gramov besed. Katere besede vsebujejo n-grami, določa slovar. Vse besede zunaj slovarja se preslikajo v simbol OOV (angl. Out-Of-Vocabulary). To pomeni, da bo beseda, ki ni v slovarju, napačno razpoznana. Napačno razpoznana beseda pa vpliva tudi na razpoznavanje besed, ki ji sledijo, saj predstavlja njihov kontekst. Pomembna je tudi velikost slovarja, saj je z velikostjo neposredno povezana kompleksnost razpoznavalnika in s kompleksnostjo tudi hitrost razpoznavanja. V sistemih razpoznavanja visoko pregibnih jezikov so neizogibni veliki slovarji, razen če je razpoznavanje omejeno na zelo specifično domeno (npr. razpoznavanje vremenske napovedi).

Beseda je praviloma osnovna enota v slovarju. Za modeliranje pregibnih jezikov so bile izvedene številne raziskave uporabe manjših osnovnih enot (morfemov, osnov in končnic besed ipd.), ki pa se niso izkazale kot bistveno boljše, saj je napovedna moč jezikovnih modelov s prehodom na manjše osnovne enote oslabljena (Sepesy Maučec idr., 2009).

2.3 Iskalni algoritmi

Naloga razpoznavalnika govora je poiskati najbolj verjetni niz besed za zajeti vhodni govor. Iskanje izvedemo s pomočjo iskalnih algoritmov (Aubert, 2002). Pri iskanju najbolj verjetnega zaporedja besed ni moč pregledati celotnega iskalnega prostora, ga pa omejujemo z različnimi hevrističnimi metodami. Razlikujemo statično omejevanje (npr. drevesna predstavitev slovarja) in dinamično omejevanje iskalnega prostora (npr. snopovno omejevanje, pogled naprej v jezikovni model ipd). Same iskalne algoritme delimo na časovno sinhrone in asinhrone glede na to, ali hipoteze v iskalnem prostoru ocenjujemo vzporedno od začetka do konca govornega segmenta ali pa vse ocenjujemo ob koncu segmentov.

Poznamo tudi dvoprehodne algoritme (Lee, Kawahara & Doshita, 1998), ki predstavljajo eno od metod za izboljšanje hitrosti delovanja algoritmov. Pri teh algoritmih najprej uporabimo samo določene jezikovne vire za samodejno razpoznavanje segmenta govora. To imenujemo prvi prehod. Kot njegov rezultat dobimo ali seznam najboljših hipotez (običajno od 100 do 1000) ali pa besedno mrežo. V drugem prehodu nato uporabimo vse razpoložljive vire in modele za ocenjevanje hipotez v seznamu oz. mreži.

2.4 Vrednotenje uspešnosti razpoznavalnika

Predlagane metode in algoritme na področju ASR najpogosteje vrednotimo posredno z uporabo rezultatov razpoznavanja govora. Vrednotenje praviloma izvajamo z ločenim testnim naborom posnetkov, ki je sicer po svojih lastnostih podoben učnemu setu, vendar ni bil uporabljen nikjer v postopku učenja akustičnih modelov. Tako je eden izmed ključnih vidikov učenja akustičnih modelov skrb, da ne pride do efekta »prenaučenja«, s čimer bi se zmanjšala njihova splošnost, nujno potrebna za uspešno vrednotenje.

Pri vrednotenju rezultatov ASR je treba upoštevati tako delež pravilno razpoznanih besed, kot tudi tiste besede, ki so bile vrinjene. Tako lahko definiramo pravilnost razpoznanih besed (ACC) kot:

$$ACC = \frac{H - I}{N}\ 100\ \%$$

pri čemer je H število vseh pravilno razpoznanih besed, I število vrinjenih besed in N število vseh besed v testni množici.

3 RAZPOZNAVANJE SLOVENSKEGA JEZIKA

Za jezikovno modeliranje je skoraj idealna angleščina. Ima malo besednih oblik in vnaprej določen vrstni red besed v povedih. Slovenščina je za razpoznavanje eden od zahtevnejših jezikov. Težave povzroča predvsem bogato pregibanje besed in relativno sproščen vrstni red, izrazit predvsem v spontanem govoru. Bogato pregibanje besed se odraža na velikosti slovarja. Za zadovoljivo pokritost besedišča mora slovar vsebovati več kot 200.000 besed, saj pomeni vsaka besedna oblika nov vnos v slovar. Po drugi strani je za učenje jezikovnega modela s tako velikim slovarjem potreben večji učni korpus, saj imamo pri majhnih korpusih težave zaradi prevelike razpršenosti podatkov. Velikost učnega korpusa danes ni več tako pereča, saj obstajajo zelo obsežne besedilne zbirke (Arhar & Gorjanc, 2007). Opozoriti pa velja, da so to zbirke pisanega jezika, ki ne odražajo značilnosti govorjenega jezika.

Razpršenost podatkov lahko zmanjšamo z lematizacijo. Lematizacija je določanje osnovne slovarske

oblike posameznim besedam v korpusu. Slovarski obliki pravimo lema. Slovar lem je v primerjavi s slovarjem besednih oblik nekajkrat manjši. Seveda pa jezikovnega modela besednih oblik ne moremo preprosto zamenjati z jezikovnim modelom lem, saj je za razpoznavalnik pomembna besedna oblika in ne zgolj lema. Uveljavilo se je modeliranje, ki razen lem modelira tudi t. i. oblikovno skladenjske oznake (angl. Morpho-Syntactic Description tags – MSD), ki če so pripete lemi, enolično določajo besedno obliko. Ker se izbrana lema lahko pojavi v mnogo različnih besednih oblikah, je število različnih MSD oznak za slovenski jezik nekajkrat večje kot za angleški jezik.

3.1 Govorni viri

Govorni in jezikovni viri so ključni pogoj za razvoj samodejnega razpoznavalnika govora. Pri tem je bistvenega pomena jezikovna odvisnost virov, saj v normalnih scenarijih razvoja samodejnega razpoznavalnika govora ne moremo uporabljati virov drugega jezika. Izdelava novega vira je časovno, stroškovno in organizacijsko zelo zahteven proces, saj je treba ročno izdelati transkripcije (prepise) z dobesednim zapisom izgovorjenega, označiti govorce, meje med segmenti, akustično ozadje itn. V povprečju je treba za izdelavo ure transkribirane govorne baze opraviti približno trideset ur dela. Navedene omejitve pri izgradnji govornih virov so še posebno izrazite pri jezikih z manjšim številom govorcev, pri čemer je manjši tudi komercialni interes. Zaradi specifičnih lastnosti jezikov virov ne moremo neposredno primerjati med seboj, temveč je treba pri primerjavi upoštevati jezikovno specifično komponento.

Slovenski jezik spada v skupino jezikov z izdelanimi osnovnimi viri za gradnjo samodejnih razpoznavalnikov govora (Kačič, 2002; Žganec Gros, Mihelič & Dobrišek, 2003). Začetki razvoja govornih virov za slovenski jezik segajo v devetdeseta leta prejšnjega stoletja. Prvi slovenski govorni viri so spadali v kategorijo razpoznavanja izoliranih in vezanih besed v telefonskem ali studijskem okolju. Na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru so bile tako razvite govorne baze SNABI, Slovenian 1000 FDB SpeechDat(II) (Kačič & Kaiser, 1998) in Polidat (Žgank, Kačič & Horvat, 2002). S stališča razvoja samodejnih razpoznavalnikov govora sta še posebno pomembni bazi SpeechDat(II) in Polidat, saj spadata v družino mednarodnih standardiziranih govornih baz, ki omogočajo razvoj govorno vodenih telekomunikacijskih storitev. Na Fakulteti za elektrotehnoiko Univerze v Ljubljani je bila za razvoj samodejnih razpoznavalnikov govora razvita baza Gopolis (Mihelič, Žganec Gros, Dobrišek, Žibert & Pavešić, 2003), ki je bila v kombinaciji z dodatnima bazama uporabljena za razvoj razpoznavalnika govora za omejeno domeno (Dobrišek, Vesnicer, Žganec Gros & Mihelič, 2006).

S stališča ASR je bistveno kompleksnejši problem razpoznavanje tekočega govora neodvisnega govorca z velikim slovarjem besed. Prva slovenska govorna baza, ki je podpirala to kategorijo govora, je bila baza Slovenian BNSI Broadcast News (Žgank, Verdonik, Zögling Markuš & Kačič, 2005), razvita leta 2005 v sodelovanju med Fakulteto za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in RTV Slovenija. Govorna baza je dostopna prek mednarodne organizacije ELRA/ELDA. Namenjena je samodejnemu razpoznavanju tekočega slovenskega govora v različnih televizijskih oddajah. To bazo smo uporabili tudi v okviru eksperimentov, predstavljenih v tem članku. Na Fakulteti za elektrotehniko Univerze v Ljubljani je bila razvita baza SiBN Broadcast News (Žibert & Mihelič, 2004), ki je prav tako namenjena razpoznavanju tekočega govora v televizijskih oddajah. V okviru sodelovanja med Fakulteto za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in državnim zborom Republike Slovenije je bila razvita govora baza SloParl (Žgank, Rotovnik, Grašič, Kos, Vlaj & Kačič, 2006), ki vsebuje posnetke sej državnega zbora. Baza obsega sto ur govora in je tako trenutno najobsežnejši govorni vir za slovenski jezik. Od preostalih slovenskih govornih baz se loči po transkripcijah govora, saj so bile transkripcije narejene na podlagi magnetogramov in ne vsebujejo dobesednega zapisa izgovorjenega. Takšno govorno bazo uporabljamo v posebnih postopkih učenja akustičnih modelov, pri čemer upoštevamo prisotnost napak v učnih transkripcijah.

Govorni bazi Slovenian BNSI Broadcast News in SloParl vsebujeta tudi besedilni korpus za učenje jezikovnih modelov samodejnega razpoznavalnika govora. Oba besedilna korpusa sta po svojih značilnostih identična govoru v govorni bazi. Tako lahko besedilna korpusa uporabljamo za izdelavo interpoliranih jezikovnih modelov, ki uspešno modelirajo tudi značilnosti govorjenega jezika. Jezikovni modeli so zaradi potrebe po dovolj velikem učnem vzorcu (reda 100 M besed) običajno izdelani na besedilnih

korpusih pisanega jezika (časopisi, knjige, splet), ki po svojih značilnostih bistveno odstopa od govorjenega jezika.

Slovenski govorni viri sicer pokrivajo osnovna področja razvoja samodejnih razpoznavalnikov govora, vendar je obseg razpoložljivih slovenskih govornih virov manjši v primerjavi z jeziki z večjim številom govorcev (angleščina, nemščina, španščina, kitajščina). Hkrati pa je slovenski jezik zaradi svojih značilnosti za ASR bistveno kompleksnejši problem. Glavni značilnosti slovenščine, ki otežita razpoznavanje govora, sta visoka pregibnost in relativno prosti vrstni red besed v stavku. Glede na izvedene analize bi tako za slovenski jezik potrebovali vsaj desetkrat večje govorne vire kot za angleški jezik (Rotovnik, Sepesy Maučec & Kačič, 2007). Če je stanje na področju osnovnih slovenskih govornih virov zadovoljivo, pa za slovenski jezik ne obstajajo bolj specifični govorni viri, ki jih poznamo za jezike z večjim številom govorcev. V to kategorijo spadajo npr. govorni viri, posneti v avtomobilu ali na motorju, govorni viri, posneti v različnih šumnih okoljih, govorni viri, posneti na sestankih, govorni viri, posneti v inteligentnem okolju itn.

V predstavljenih eksperimentih smo uporabili govorno bazo Slovenian BNSI Broadcast News. Baza vsebuje transkribirane posnetke 42 dnevnoinformativnih oddaj RTV Slovenija (TV Dnevnik, Odmevi) iz obdobja 1999–2003. Kot učni korpus uporabljamo trideset ur posnetkov, tri ure so namenjene razvojnemu testiranju ter tri ure vrednotenju. Posnetki vsebujejo 1565 različnih govorcev, od tega 1069 moških in 477 žensk. Za 19 govorcev ni bilo mogoče zanesljivo določiti spola zaradi značilnosti akustičnega kanala (kratki odseki, prekrivajoči se govori). Za vsakega govorca je bilo ustrezno določeno njegovo narečje. V transkripcijah so ustrezno označene akustične lastnosti (studio/telefon, akustično ozadje) posnetkov ter lastnosti govora in govorcev (brani/spontani govor, prekrivanje govorcev, tuji govorci). Na podlagi teh lastnosti so segmenti razdeljeni v ustrezne »f-kategorije«. Glede na vsebino prispevka so bili posnetki razdeljeni v petnajst različnih topikov, s pomočjo katerih je mogoče omejiti domeno samodejnega razpoznavalnika govora in tako izboljšati rezultate. V transkripcijah baze BNSI je 268.000 besed, od tega 37.000 različnih.

3.2 Jezikovni viri

Za izdelavo jezikovnih modelov potrebujemo dovolj velike korpuse jezika, ki nam služijo kot učna množica. Prvi obsežen korpus slovenskega jezika je bil korpus FIDA, ki se je kasneje nadgradil v korpus FidaPLUS (Arhar & Gorjanc, 2007), ki ga tudi uporabljamo za gradnjo jezikovnih modelov v razpoznavalniku UMB Broadcast News. FidaPLUS je največji korpus, ki nam je trenutno na voljo. Vsebuje približno 621 milijonov besed. Največji delež besedil glede na zvrst predstavljajo neumetnostna nestrokovna besedila. Glede na tip prevladujeta časopisno in revijalno gradivo. Podrobnejše podatke o sestavljenosti korpusa lahko najdemo v Arhar & Gorjanc (2007). Besede v korpusu so tudi samodejno označene s pripadajočimi lemami in oznakami MSD.

Korpus FidaPLUS je bil kasneje nadgrajen še v korpus Gigafida (Arhar Holdt, Kosem & Logar Berginc, 2012), ki nam trenutno še ni na voljo. Ta korpus vsebuje približno 1,1 milijarde besed, ki so prav tako označene z lemami in oznakami MSD.

Za razpoznavanje govora so se poleg osnovnih besednih oblik izkazale kot uporabne tudi dodatne jezikovne informacije. Za slovenski jezik so tukaj lahko uporabne besedne leme in oznake MSD. Da jih lahko uporabimo v razpoznavanju govora, potrebujemo jezikovne vire s čim bolj natančnimi oznakami in pomoč označevalnika med samim postopkom razpoznavanja.

Ker vsako samodejno označevanje korpusov z oznakami MSD vnaša napake, je smiselno uporabiti korpuse, ki so bili označeni ali vsaj pregledani ročno. Tak korpus je npr. jos100k (Erjavec & Krek, 2008), ki je nastal v okviru projekta Jezikovno označevanje slovenščine (JOS). Korpus je bil kasneje v projektu Sporazumevanje v slovenskem jeziku (SSJ) razširjen v korpus ssj500k (Arhar, 2009). Ta vsebuje približno 500.000 besed, označenih z oznakami MSD, ki so pregledane ročno.

Ta korpus je sicer veliko manjši od korpusa FidaPLUS, vendar je kljub temu uporaben za izdelovanje statističnih modelov oznak MSD. Medtem ko slovarji besed lahko vsebujejo do več sto tisoč enot, lahko vsebujejo slovarji oznak MSD le nekaj sto do nekaj tisoč enot, odvisno od kompleksnosti oznak. V okvirju projekta JOS so bila definirana tudi pravila za obliko oznak MSD. Po sistemu JOS poznamo skupaj 1.903 različnih oznak MSD. Število teh oznak lahko zmanjšamo s poenostavljanjem. Tako lahko iz oznak izpu-

ščamo podatke, ki so manj pomembni za razpoznavanje. Zaradi veliko manjšega števila različnih enot v slovarju je treba za gradnjo statističnega modela oceniti bistveno manj parametrov. Zato za gradnjo modelov oznak MSD ni potrebna tako velika učna množica kot pri modelih besed.

Prav tako je v okviru projekta SSJ nastal oblikoskladenjski označevalnik in lematizator Obeliks (Grčar, Krek & Dobrovoljc, 2012). Označevalnik prav tako potrebuje statistične modele, ki so naučeni na neki učni množici. Označevalnik pripisuje besedam leme in oznake MSD po sistemu JOS.

4 EKSPERIMENTALNI SISTEM

Vsi predstavljeni eksperimenti so bili izvedeni na razpoznavalniku tekočega govora UMB Broadcast News (Žgank & Sepesy Maučec, 2010). Trenutno v njem uporabljamo dvoprehodni algoritem razpoznavanja. Za učenje akustičnih modelov in razpoznavanje v prvem prehodu smo uporabljali orodja iz zbirke HTK (Young, Jansen, Odell, Ollason & Woodland, 1996), za gradnjo slovarjev, jezikovnih modelov in razpoznavanje v drugem prehodu pa orodja iz zbirke SRILM (Stolcke, Zheng, Wang & Abrash, 2011).

Prvi korak v postopku akustičnega modeliranja je izločanje značilk iz govornega signala. Vhodni signal s funkcijo okna dolžine 25 ms, ki ga premikamo s koraki10 ms, razdelimo na kratkočasovne vzorce. Po izvedbi predpoudarjanja izračunamo 12 mel-kepstralnih koeficientov in energijo ter njihove prve in druge odvode. Končni vektor značilk ima tako 39 elementov.

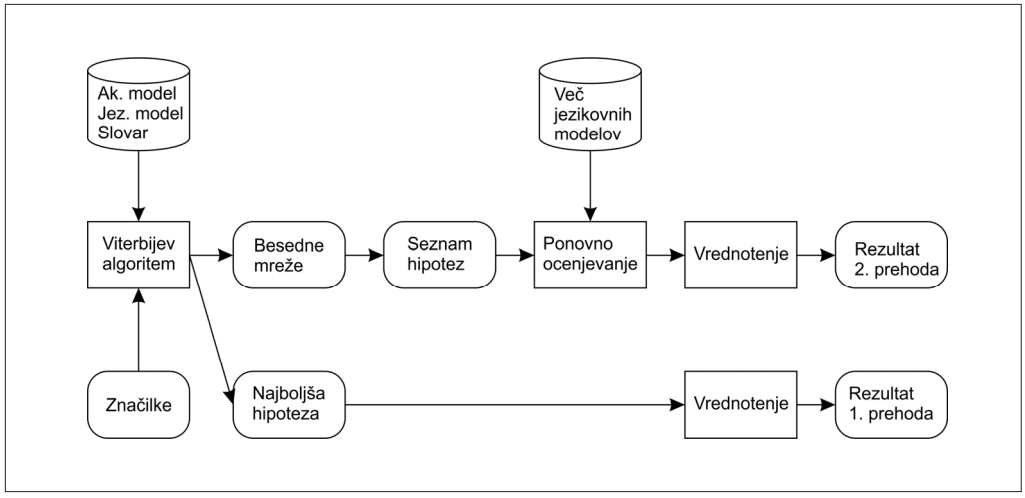
Postopek učenja akustičnih modelov poteka v treh korakih, pri čemer se postopoma izboljšuje kakovost akustičnih modelov. Kot osnovno akustično enoto smo uporabili grafeme, saj so predhodne analize pokazale, da je tako mogoče učiti kakovostne akustične modele (Žgank & Sepesy Maučec, 2010). V nadaljevanju bomo za akustične modele uporabljali poimenovanje fonem in trifon, kljub temu da je bila osnovna akustična enota grafem. V učnem setu smo uporabili 24 oddaj. V prvem koraku izvedemo inicializacijo parametrov akustičnih modelov z globalnimi vrednostmi. Temu sledi več ponovitev učnega Baum-Welchevega algoritma. S tako naučenimi akustičnimi modeli izvedemo prisilno poravnavo transkripcij, s katero se izboljša njihova kakovost. Sledi drugi korak s ponovim učenjem akustičnih modelov od začetka, vendar tokrat z izboljšanimi transkripcijami. Inicializacija vrednosti parametrov prikritih modelov Markova se tokrat izvrši ločeno za vsak fonem posebej.

Akustični modeli, naučeni v drugem koraku, služijo za izhodišče tretjega koraka, v katerem se najprej tvorijo kontekstno odvisni akustični modeli – trifoni, pri katerih upoštevamo predhodni in naslednji fonem. Posledično zelo naraste število prostih parametrov akustičnih modelov, ki jih je treba oceniti med postopkom učenja. Zato uporabimo postopek združevanja z odločitvenim drevesom, pri čemer na podlagi podatkovne metrike združimo stanja oz. celotne modele, ki so med seboj dovolj podobni. Odločitveno drevo zgradimo na podlagi fonetičnih razredov, ki so bili v predstavljenem eksperimentu tvorjeni s podatkovno vodeno metodo na podlagi matrike zamenjav fonemov. Akustični modeli, združeni z odločitvenim drevesom, so bili izhodišče za zadnji korak učenja, v katerem se je število Gaussovih porazdelitvenih funkcij verjetnosti korakoma povečalo do 16 na stanje. Takšni akustični modeli so bili uporabljeni za vrednotenje samodejnega razpoznavalnika govora.

Pred razpoznavanjem govora smo zgradili vrsto jezikovnih modelov, ki smo jih primerjali glede na uspešnost v razpoznavalniku. Tako smo najprej definirali različne velikosti slovarjev od 60.000 (60 k) do 300.000 (300 k) besed. Preizkušali smo dva načina gradnje slovarjev. V prvem načinu (FP) smo slovar gradili tako, da smo mu dodajali besede v vrstnem redu, ki ga je določala njihova pogostost v korpusu FidaPLUS. Ko smo dosegli želeno velikost slovarja, smo v slovar dodali še vse besede, ki so se pojavile z enako frekvenco kot nazadnje dodana beseda. V drugem načinu gradnje slovarja (BNSI+FP) smo najprej v slovar vključili vse besede iz govorne učne množice BNSI, nato smo dodajali besede iz besedilnega korpusa BNSI (iNews) in nazadnje besede iz korpusa FidaPLUS.

Pred gradnjo jezikovnih modelov smo pogledali deleže besed zunaj slovarja, ki se pojavijo na testni množici BNSI glede na oba načina gradnje slovarja. Po pregledu rezultatov smo se odločili, da bomo jezikovne modele gradili le na slovarjih, sestavljenih po prvem načinu (FP).

Nato smo zgradili standardne bigramske, trigramske in štirigramske modele. Pri tem smo uporabljali tako glajenje Good-Turing kot Knesser-Ney. Raziskali smo tudi vpliv velikosti učne množice, zato smo kot učno množico enkrat uporabili celotni korpus FidaPLUS, drugič pa le njegov del – približno devet odstotkov.



Slika 3: **Blokovna shema poteka razpoznavanja**

Splošna shema našega eksperimentalnega sistema je podana na sliki 3. Iskalni algoritem v prvem prehodu je sinhroni Viterbijev algoritem s snopovnim omejevanjem, ki je implementiran v orodju HDecode. Za vsak vhodni akustični segment nam algoritem vrne najboljšo hipotezo in besedno mrežo, ki pomeni iskalni prostor algoritma ob koncu segmenta. Najboljšo hipotezo določimo po uteženem razmerju med verjetnostima, dobljenima z akustičnim in jezikovnim modelom. Za določitev optimalnih vrednosti uteži smo uporabili rezultate razpoznavanja na razvojni množici BNSI.

Kadar neposredno vrednotimo uspešnost razpoznavanja na najboljši hipotezi, dobimo rezultate prvega prehoda. Na podlagi teh rezultatov smo se odločili, katere sisteme prvega prehoda (glede na različne jezikovne modele) bomo uporabili v dvoprehodnem algoritmu.

Pred drugim prehodom razpoznavanja besedne mreže pretvorimo v sezname sto najboljših hipotez, ki jih lahko razberemo iz njih. V nekaterih segmentih je to število tudi manjše, ker ni mogoče tvoriti takšnega števila hipotez. Hipoteze nato oblikoskladenjsko označimo z označevalnikom Obeliks. V naslednjem koraku oznake poenostavimo tako, da vsebujejo le podatek o besedni vrsti, spolu, sklonu, številu in osebi razpoznane besede.

V drugem prehodu hipoteze ponovno ovrednotimo z novimi jezikovnimi modeli. Teh modelov je sedaj lahko tudi več. Podobno kot pri prvem prehodu utežimo verjetnosti, dobljene s posameznimi modeli. Pri tem je treba ponovno uporabiti razvojno množico za iskanje optimalnih vrednosti uteži. Kot končni rezultat algoritem vrne hipotezo, ki ima po drugem prehodu največjo verjetnost.

Za vrednotenje označenih hipotez v drugem prehodu smo zgradili modele oznak MSD. Kot učno množico smo uporabili korpus ssj500k, v katerem smo oznake poenostavili na enak način kot v označenih hipotezah razpoznavalnika.

5 REZULTATI

V Donaj & Kačič (2012) smo že predstavili vpliv velikosti slovarja na delež besed OOV na testni množici BNSI. Tam uporabljeni slovarji so bili grajeni le glede na korpus FidaPLUS. Tabela 2 podaja k temu še rezultate OOV, kadar gradimo slovarje enakih velikosti po drugem načinu (BNSI + FP).

Tabela 2: **Delež besed OOV glede na način gradnje slovarja in njegovo velikost**

Velikost slovarja	Prvi način (FP)	Drugi način (BNSI + FP)
60 k	6,94	5,09
100 k	3,44	3,23
200 k	1,64	2,08
300 k	1,02	1,44

Iz rezultatov vidimo, je pri manjših velikostih slovarja bolj ugodno upoštevati najprej tekstovni korpus BNSI, pri večjih slovarjih pa je položaj ravno nasproten. Manjši delež besed zunaj slovarja dobimo, ko uporabljamo samo korpus FidaPLUS. Vzrok za to vidimo v dejstvu, da se pri drugem načinu gradnje v slovar vključijo besede, ki se v učni množici in v

besedilnem delu BNSI pojavijo zelo redko, medtem ko se ne vključijo besede iz korpusa FidaPLUS, ki se v testni množici pojavijo pogosteje.

V tabeli 3 so predstavljeni rezultati razpoznavanja prvega prehoda pri različnih velikostih učne množice, različnih velikostih slovarja in pri uporabi bigramskih (2 g) in trigramskih (3 g) jezikovnih modelov. V tabeli 4 so podani tudi faktorji realnega časa, s katerimi je potekalo razpoznavanje v teh primerih.

Tabela 3: **Uspešnost razpoznavanja glede na velikost učne množice in jezikovni model**

Slovar	Red modela	9 % Fidaplus	100 % Fidaplus
60 k	2 g	64,05	66,09
60 k	3 g	65,80	69,23
300 k	2 g	68,11	70,77
300 k	3 g	69,90	74,33

Tabela 4: **Faktorji realnega časa pri razpoznavanju glede velikost učne množice in jezikovni model**

Slovar	Red modela	9 % Fidaplus	100 % Fidaplus
60 k	2 g	6,04	6,30
60 k	3 g	8,58	18,46
300 k	2 g	13,35	12,66
300 k	3 g	19,16	37,09

Iz podatkov v tabeli 3 lahko vidimo, da se pri povečanju učne množice, povečanju slovarja in povečanju reda modela opazno izboljša uspešnost razpoznavanja. Izboljšanje uspešnosti ob povečanju velikosti slovarja je v vseh primerih približno 4 do 5 odstotkov, kar je v velikostnem redu zmanjšanja besed OOV pri spremembi velikosti slovarja. Spremembe v uspešnosti ob povečanju reda modela iz bigramskega na trigramskega so odvisne od velikosti učne množice. Medtem ko sta pri uporabi manjše učne množice spremembi 1,75 in 1,79 odstotka, sta pri uporabi večje učne množice spremembi 3,14 in 3,56 odstotka. Iz podatkov v tabeli 4 je razvidno, da tako povečanje slovarja kot tudi zvišanje reda modela poveča časovno zahtevnost razpoznavanja govora. Pri povečanju slovarja se faktor realnega časa poveča za približno 2. Pri zvišanju reda modela pa je ta faktor različen glede na velikost učnega korpusa. V primeru uporabe celotnega korpusa se velikost faktorja poveča približno za 3. Pri uporabi manjšega korpusa je povečanje veliko manjše.

Na podlagi teh podatkov lahko sklepamo, da bi dodatno povečanje učne množice (npr. z uporabo korpusa Gidafida) še dodatno povečalo uspešnost razpoznavanja, ki bo bolj izrazito pri uporabi trigramskega modela.

V tabeli 5 so prikazani rezultati uspešnost razpoznavanja pri uporabi modelov z modificiranim glajenjem Knesser-Ney, ki sta ga predstavila Chen & Goodman (1999) in razlika v uspešnosti glede na ustrezni model z glajenjem Good-Turing.

Tabela 5: **Uspešnost razpoznavanja z modificiranim glajenjem Knesser-Ney**

Slovar	Red modela	Acc (KN)	Acc (KN) – Acc (GT)
60 k	2 g	66,15	+ 0,06
60 k	3 g	69,04	+ 0,19
300 k	2 g	70,71	– 0,06
300 k	3 g	74,12	– 0,21

Iz rezultatov vidimo, da so modeli z modificiranim glajenjem Knesser-Ney uspešnejši le pri manjših slovarjih, medtem ko so pri večjih slovarjih uspešnejši modeli z glajenjem Good-Turing. V obeh primerih so razlike le majhne.

V vseh poskusih smo dobili besedne mreže, s katerimi bi lahko nadaljevali razpoznavanje v drugem prehodu, vendar smo se omejili le na rezultate, ki smo jih dobili pri slovarju 300 k in glajenjem GT. Prva različica tega algoritma je bila predstavljena v Donaj & Kačič (2012). Pokazano je bilo, da lahko z uporabo dvoprehodnega dosežemo primerljive uspešnosti ob bistveno krajšem času razpoznavanja. Prav tako je bilo pokazano, da uporabi trigramskih in štirigramskih modelov v drugem prehodu dajeta enake rezultate.

Za vrednotenje hipotez v drugem prehodu smo uporabili dva jezikovna modela. Prvi je standardni besedni trigramski model, drugi pa je trigramski model oznak MSD. V tabeli 6 so predstavljeni rezultati dvoprehodnega algoritma za istočasno vrednotenje z trigramskim modelom besed in trigramskim modelom oznak MSD.

Tabela 6: **Rezultati v dvoprehodnem algoritmu**

Prvi prehod	74,33 %
Drugi prehod	74,85 %
Sprememba	0,52 %

Iz podatkov vidimo, da smo lahko s pomočjo preprostega modela oznak MSD izboljšali uspešnost razpoznavanja za 0,52 odstotka.

6 SKLEP

V prispevku smo predstavili osnovne pojme s področja samodejnega razpoznavanja govora in govorne ter jezikovne vire za slovenščino, ki jih uporabljamo na tem področju. Razpoznavanji tekočega in spontanega govora sta nalogi z veliko prostora za vpeljevanje izboljšav tako v akustičnem kot v jezikovnem modeliranju. Predstavljeni rezultati kažejo na pomembnost ustreznih jezikovnih virov. Tukaj sta pomembna tako obseg virov kot tudi njihova dodatno obogatena vsebina, kot sta lematizacija in oblikoskladenjsko označevanje besedila.

Prestavljeni rezultati uporabe oblikoskladenjskih oznak v jezikovnem modeliranju pomenijo le začetek dela na tem področju. Zaradi svoje kompleksnosti v kombinaciji z uveljavljenimi jezikovnimi modeli ponujajo ti modeli veliko možnosti za teoretične in praktične raziskave.

Naše nadaljnje raziskave na področju ASR bodo usmerjene tudi v uporabo novih virov za izdelavo modelov, kot sta npr. korpusa Gigafida in GOS, kot tudi na izboljšano uporabo razpoložljivih informacij v korpusih.

Medtem ko je samodejno razpoznavanje govora že uporabno v omejenih domenah z majhnimi slovarji besed, pa trenutni rezultati razpoznavanja tekočega govora z velikim slovarjem besed še niso zadovoljivi za praktične aplikacije. Zato bodo še potrebne raziskave, ki bodo usmerjene tako v izboljšanje uspešnosti kot tudi hitrosti razpoznavanja govora. Zaradi težavnosti razpoznavanja slovenskega govora bo potrebno tudi nadaljnje delo na področju izdelave jezikovnih virov slovenščine. Le s takšnim celovitim pristopom bomo lahko zagotovili stik našega jezika s sodobnimi trendi v informacijsko-komunikacijskih tehnologijah.

LITERATURA

[1] Arhar, Š. & Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), 95–110.

[2] Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54(3–4), 43–56.

[3] Arhar Holdt, Š., Kosem, I. & Logar Berginc, N. (2012). Izdelava korpusa Gigafida in njegovega spletnega vmesnika. *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenija, 16–21.

[4] Aubert, X. L. (2002). An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer speech & language*, 16(1), 89–114.

[5] Chen, S. F. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer speech & language*, 13(4), 359–393.

[6] Dobrišek, S., Vesnice, B., Žganec Gros, J. & Mihelič, F. (2006). Uporaba kanoničega govornega akustičnega modela za prilagajanje prostora govornih akustičnih značilk. *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba*, Ljubljana, Slovenija, 89–92.

[7] Donaj, G. & Kačič, Z. (2012). Širjenje slovarja in dvoprehodni algoritem v razpoznavalniku tekočega govora UMB Broadcast News. *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenija, 48–51.

[8] Erjavec, T. & Krek, S. (2008). Oblikoskladenjske specifikacije in označeni korpusi JOS. *Zbornik Šeste konference Jezikovne tehnologije*, Ljubljana, Slovenija, 49–53.

[9] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4), 237–264.

[10] Grčar, M., Krek, S. & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenija, 89–94.

[11] Kačič, Z. & Kaiser, J. (1998). Development of Slovenian SpeechDat database. *First International Conference on Language Resources and Evaluation, Workshop on speech database development for Central and Eastern European languages,* Granada, Spain.

[12] Kačič, Z. (2002). Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. *Jezikovne tehnologije: zbornik konference*, Ljubljana, Slovenija, 111–115.

[13] Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on acoustics, speech and signal processing*, 35(3), 400– 401.

[14] Lee, A., Kawahara, T. & Doshita, S. (1998). An efficient two--pass search algorithm using word trellis index. *Proceeding of the 5th International Conference on Spoken Language Processing*, Sydney, Australia.

[15] Mihelič, F., Žganec 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.

[16] Rotovnik, T., Sepesy Maučec, M. & Kačič, Z. (2007). Large vocabulary continuous speech recognition of an Inflected language using stems and endings. *Speech communication*, 49(6), 437–452.

[17] Sepesy Maučec, M., Rotovnik, T., Kačič, Z. & Brest, J. (2009). Using data-driven subword units in language model of highly inflective Slovenian language. *International journal of pattern recognition artificial intelligence*, 23(2), 287–312.

[18] Stolcke, A., Zheng, J., Wang, W. & Abrash, V. (2011). SRILM at sixteen: Update and outlook. *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*.

[19] Young, S., Jansen, J., Odell, J., Ollason, D. & Woodland, P. (1996). *The HTK book*. Cambridge University.

[20] Žganec Gros, J., Mihelič, F. & Dobrišek, S. (2003). Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo*, 48(3–4), 47–59.

[21] Žgank, A., Kačič, Z. & Horvat, B. (2002). Preliminary evaluation of Slovenian mobile database PoliDat. *Third international conference on language resources and evaluation*, Las Palmas de Grand Canaria, Spain, 564–568.
[22] Žgank, A., Rotovnik, T., Grašič, M., Kos, M., Vlaj, D. & Kačič, Z. (2006). SloParl – Slovenian parlamentary speech and text corpus for large vocabulary continuous speech recognition. *Ninth international conference on spoken language processing*, Pittsburgh, PA, USA, 197–200.
[23] Žgank, A., Rotovnik, T. & Sepesy Maučec, M. (2008). Slovenian spontaneous speech recognition and acoustic modeling of filled pauses and onomatopoeas. *WSEAS transactions on signal processing,* 4(7), 388–39.
[24] Žgank, A., Sepesy Maučec, M. (2010). Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Zbornik Sedme konference Jezikovne tehnologije*, Ljubljana, Slovenija, 28–31.
[25] Žgank, A., Verdonik, D., Zögling Markuš, A. & Kačič, Z. (2005). BNSI Slovenian broadcast news database – speech and text corpus. *9th European conference on speech communication and technology*, Lisbon, Portugal, 1537–1540.
[26] Žibert, J. & Mihelič, F. (2004). Development of Slovenian broadcast news speech database. *Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2095–2098.

▪

Gregor Donaj je diplomiral iz elektrotehnike na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in iz matematike na Fakulteti za naravoslovje in matematiko Univerze v Mariboru. Trenutno je doktorski študent in zaposlen kot mladi raziskovalec na Fakulteti za elektrotehniko, računalništvo in informatiko. Raziskovalno se ukvarja z jezikovnim modeliranjem za avtomatsko razpoznavanje govora.

▪

Andrej Žgank je leta 2003 doktoriral na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Na tej fakulteti je tudi zaposlen kot izredni profesor za področje telekomunikacije. Njegovo raziskovalno področje obsega večjezičnost, križnojezično razpoznavanje govora, akustično modeliranje pri razpoznavalniku govora z velikim slovarjem in gradnja jezikovnih virov.

▪

Mirjam Sepesy Maučec je izredna profesorica za področje telekomunikacije na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Ob pedagoškem delu je raziskovalno aktivna v številnih nacionalnih in mednarodnih projektih s področja jezikovnih tehnologij. Njeno raziskovalno področje obsega statistično jezikovno modeliranje in strojno prevajanje.