Razpoznavalnik tekočega govora UMB Broadcast News 2020: prehod na arhitekturo z nevronskimi mrežami

Andrej Žgank, Mirjam Sepesy Maučec, Gregor Donaj

Laboratorij za digitalno procesiranje signalov, Fakulteta za elektrotehniko, računalništvo in informatiko,
Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
andrej.zgank@um.si, mirjam.sepesy@um.si, gregor.donaj@um.si

Povzetek

V članku bomo predstavili novo verzijo slovenskega avtomatskega razpoznavalnika govora za domeno televizijskih oddaj UMB Broadcast News 2020. Prehod na arhitekturo globokih nevronskih mrež ob upoštevanju razpoložljivih govornih virov je osrednja točka nove verzije sistema. Za učenje razpoznavalnika govora smo uporabili bazi UMB BNSI Broadcast News in IETK-TV. Skupni obseg govornih posnetkov je znašal 66 ur. Vzporedno z globokimi nevronskimi mrežami smo povečali slovar razpoznavanja govora, ki je tako znašal 250k besed. Na ta način smo znižali delež besed izven slovarja na 1,33 %. Z razpoznavanjem govora na testni množici smo dosegli najboljši WER 17,19 %, kar predstavlja 9,62 % izboljšanje glede na predhodno verzijo sistema. Med procesom vrednotenja smo izvedli tudi podrobnejšo analizo napak razpoznavanja govora na osnovi lem in F-razredov.

UMB Broadcast News 2020 continuous speech recognition system: a new approach based on neural networks

This paper presents a new version of the Slovenian UMB Broadcast News 2020 automatic speech recognition system for the TV domain. The main contribution of the new version is the transition to deep neural networks architecture, considering available speech resources. We used the UMB BNSI Broadcast News and IETK-TV speech databases to train the speech recognizer. The total amount of recordings was 66 hours. In parallel with the deep neural networks, the vocabulary size was increased from 64k to 250k words. In this way, we reduced the out of vocabulary rate to 1.33%. Speech recognition on the test set achieved the best WER of 17.19%, which represents a 9.62% improvement over our previous version of the system. During the evaluation process, we also performed a more detailed analysis of speech recognition errors based on lemmas and F-conditions.

1. Uvod

Napredna pametna okolja v veliki meri temeljijo na zajemanju širokega nabora informacij, ki imajo lahko ali uporabniški ali medijski izvor. Razvoj tehnologij je v zadnjem desetletju privedel do skokovitega povečanja količine podatkov, ki lahko služijo kot vir informacij. Hkrati prihaja v ospredje tudi naravna interakcija z napravami s pomočjo govora, kar uporabniku olajša rokovanje z napravami in zmanjšuje digitalno ločnico. Učinkovit preplet teh razvojnih trendov predstavlja inteligentno okolje interneta stvari.

Ena izmed jedrnih tehnologij, ki omogočajo ustrezno podporo za zajemanje informacij, tako iz uporabniškega ali medijskega toka, kot tudi iz uporabniškega vmesnika, je avtomatsko razpoznavanje govora (ASR). Deluje lahko v zelo različnih scenarijih, od preprostega ukaznega krmiljenja do zahtevnih sistemov za razpoznavanje spontanega govora več govorcev. S kompleksnostjo scenarija je praviloma obratno sorazmerna uspešnost razpoznavanja govora, pomemben vidik pa predstavlja tudi računska zahtevnost, ki lahko pogosto trči ob vprašanja zagotavljanja zasebnosti govora, kadar je v uporabi procesiranje v oblaku.

Področje avtomatskega razpoznavanja govora je neločljivo povezano z razpoložljivostjo govornih virov za posamezni jezik. Tukaj nastopi težava pri jezikih, za katere obstaja manjši (komercialni) interes za implementacijo ASR, kar se lahko še dodatno potencira s posebnostmi določenih jezikov, ki otežijo razpoznavanje govora. V kategorijo za procesiranje zahtevnih jezikov sodi tudi slovenščina, za katero je značilna visoka pregibnost besed in relativno prost vrstni red besed v stavku.

Razvoj prvih sistemov govornih tehnologij za slovenščino se je začel že pred 30 leti, vendar finančno in časovno zahteven razvoj govornih virov v zadnjem desetletju ni uspel slediti svetovnim trendom. Postopki globokega učenja razpoznavalnikov govora namreč za učinkovito delovanje potrebujejo govorne baze v obsegu več 100 oz. 1000 ur transkribiranih posnetkov. Za področje slovenskega jezika pričakujemo razpoložljivost tako obsežnih govornih virov kot enega od rezultatov projekta RSDO – Razvoj slovenščine v digitalnem okolju, ki bo potekal do leta 2022.

Cilj pričujočega raziskovalnega dela[1] je predstaviti nadgradnjo našega sistema za avtomatsko razpoznavanje govora, ki deluje za domeno televizijskih oddaj. Podati želimo oceno, kakšen primanjkljaj pri prehodu na nove arhitekture predstavljajo omejene govorne baze za slovenski jezik. V tem okviru bomo izvedli tudi analizo napak razpoznavanja govora glede na značilnosti jezika, ki so pomembne za rezultate razpoznavanja govora. Delo smo zasnovali na slovenski bazi televizijskih dnevno-informativnih oddaj UMB BNSI Broadcast News (Žgank et al., 2004) in IETK-TV, saj ti dve govorni bazi trenutno še vedno predstavljata najprimernejši vir za takšno analizo, hkrati pa omogočata tudi primerljivost rezultatov s starejšimi sistemi avtomatskega razpoznavanja govora (Žgank et al., 2014).

V nadaljevanju članka bomo najprej predstavili trenutno stanje na področju govornih virov za slovenski jezik. V tretjem poglavju bo sledila kratka predstavitev

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069 in s strani Ministrstva za kulturo RS v okviru projekta RSDO – Razvoj slovenščine v digitalnem okolju.

teoretičnega ozadja pristopov, ki se danes uporabljajo pri gradnji avtomatskih razpoznavalnikov govora. V četrtem poglavju bomo predstavili govorne in jezikovne vire, ki smo jih uporabili pri raziskavi. Postopek izdelave akustičnih in jezikovnih modelov eksperimentalnega sistema bomo opisali v petem poglavju. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v šestem poglavju. V zadnjem poglavju bomo podali zaključne misli.

2. Pregled govornih virov za slovenski jezik

Že v uvodu smo zapisali, da predstavljajo govorni viri ključno komponento za razvoj avtomatskega razpoznavalnika govora. Pomembno je, da pri tem s svojimi značilnostmi in obsegom materiala vplivajo tudi na to, katero arhitekturo nevronskih mrež, ki so danes najbolj aktualna tehnologija pri razvoju razpoznavalnikov, bo možno uspešno naučiti.

Dosedanji razvoj govornih virov za slovenski jezik lahko razdelimo na dve obdobji. V prvem obdobju, ki se je začelo v devetdesetih letih prejšnjega stoletja, je bil poudarek na razvoju govornih baz za omejene scenarije izoliranih ali vezanih besed. Snemalni kanal je bil ali studio ali telefon, obseg govornega material pa praviloma med 10 in 15 ur. V to skupino lahko uvrstimo govorne baze: FDB 1000 Slovenian SpeechDat(II) (Kaiser in Kačič, 1997), Polidat (Žgank et al., 2002), Gopolis (Dobrišek et al., 1998), VNTV/VNRAD (Žibert et al., 2003) in SNABI. Delni sklopi naštetih baz že vsebujejo tudi tekoči govor, vendar je zaradi omejene količine govornega materiala praktičen razvoj splošnega razpoznavalnika govora še nemogoč.

V drugem obdobju razvoja govornih baz za slovenski jezik, ki se je začelo okoli leta 2004, se aktivnosti osredotočijo na tekoči govor. Bistveno se razširi domena vključenega materiala, kot snemalni kanal pa se dodatno pojavi televizija oziroma druge oblike javnega govora, kot so npr. predavanja. Obseg govornih baz se poveča na nekaj 10 ur posnetkov. Sem lahko prištejemo sledeče televizijske baze: UMB BNSI Broadcast News (36 ur) (Žgank et al., 2004), SiBN Broadcast News (36 ur) (Žibert in Mihelič, 2004), IETK-TV (30 ur) in GOS javni podkorpus (42 ur) (Verdonik et al., 2013). Predavanja najdemo v bazi SI TEDx-UM (54 ur, avtomatske transkripcije) (Žgank et al., 2016) in bazi GOS-Videolectures (22 ur) (Verdonik, 2018). Baza SloParl (Žgank et al., 2006) vsebuje 100 ur posnetkov in magnetogramov parlamentarnih razprav iz DZ RS, baza SOFES (Dobrišek et al., 2017) pa 10 ur posnetkov s poizvedbami po letalskih informacijah.

Dostopnost predstavljenih govornih baz pokriva skoraj celotni spekter možnosti. Nekatere so prosto dostopne preko iniciative Clarin oz. na spletnih straneh avtorjev. Druge baze so dostopne proti plačilu preko organizacije ELRA. Del baz pa je namenjen izključno interni uporabi in tako nedostopen širši raziskovalni skupnosti.

Skupna dolžina transkribiranih posnetkov v predstavljenih govornih bazah je približno 250 ur. Dodatnih 150 ur posnetkov je transkribiranih samo avtomatsko ali v obliki magnetogramov. Tudi če bi kljub različnim omejitvam v dostopnosti uspeli združiti vse govorne baze, prihaja med njimi v zasnovi do tako velikih razlik, da bi bilo učenje razpoznavalnika govora na takšen način neizvedljivo. Ob upoštevanju kriterija sorodnosti in dostopnosti govornih baz je trenutno praktično možno za učenje slovenskega razpoznavalnika govora uporabiti med 50 in 100 urami posnetkov. Takšen obseg učnega materiala je premajhen za uporabo naprednejših arhitektur globokega učenja.

To dejstvo lepo kaže na nujno potrebo po tretjem obdobju v razvoju govornih baz za slovenski jezik, kjer je cilj pridobiti nekaj 100 do 1.000 ur posnetkov, ki so prosto dostopni in omogočajo potencialno kombiniranje virov v prihodnosti. V to kategorijo bo sodila govorna baza, ki nastaja v okviru projekta RSDO.

3. Arhitekture za avtomatsko razpoznavanja govora

Na področju arhitekture avtomatskih razpoznavalnikov govora obstajata dve glavni skupini. Prvo predstavljajo sistemi s prikritimi modeli Markova, ki so bili glavni gradnik akustičnega modeliranja v preteklosti. Drugo skupino, ki je danes standardna, pa predstavljajo sistemi na osnovi nevronskih mrež.

3.1. Prikriti modeli Markova

Prikriti modeli Markova uporabljajo pristop statističnega modeliranja, kjer na osnovi vhodnih vektorjev značilk ocenjujejo verjetnost hipoteze izgovorjenega besedila. Običajno se uporabljajo večstanjski levo-desni prikriti modeli, kjer je porazdelitvena funkcija gostote verjetnosti modelirana s skupino uteženih multivariantnih Gaussovih porazdelitvenih funkcij. Z vidika računske kompleksnosti in količine zahtevanega učnega materiala gre praviloma za manj zahtevne sisteme v primerjavi z globokimi nevronskimi mrežami.

3.2. Globoke nevronske mreže

Nevronske mreže predstavljajo pristop na področju strojnega učenja, ki deloma posnema dogajanje v nevronskem sistemu. Mreže so sestavljene iz nevronov, ki so razporejeni v plasti – vhodno plast, notranje plasti in izhodno plast.

Vsak nevron izvaja matematično operacijo, kjer najprej izračuna uteženo vsoto vrednosti na njegovih vhodih, nato pa to vsoto uporabi v aktivacijski funkciji, da dobi izhodno vrednost nevrona. Izhodi nevronov so potem povezani na vhode drugih nevronov.

V zadnjih letih so nevronske mreže postale popularne na raznih področjih strojnega učenja, tudi pri razpoznavanju govora. Ker pa gre tukaj za razpoznavanje časovne vrste, niso vse arhitekture nevronskih mrež primerne.

Med korakom učenja se nevronska mreža prilagaja na učne podatke tako, da spreminja uteži. Pri uporabi pa nato dajemo nove podatke na vhodno plast omrežja ter opazujemo rezultate na izhodni plasti.

4. Govorni in jezikovni viri

Osrednji vir podatkov za modeliranje je predstavljala govorna baza UMB BNSI Broadcast News (Žgank et al., 2004), ki jo distribuira organizacija ELRA (2020). Govorna baza vsebuje posnetke dnevno-informativnih televizijskih oddaj RTV Slovenija v obsegu 36 ur. Od tega je 30 ur namenjenih učenju akustičnih modelov. Oddaje so nastale v letih 1999–2003, tako da je bila z vidika naprav uporabljenih v produkciji tehnologija delno drugačna, kot

jo srečamo danes (npr.: snemalne naprave z izgubnimi kodeki, povezave VoIP, spletne komunikacijske platforme). V bazi je skupaj 1.565 govorcev, od tega 1.069 moških in 477 žensk. Za 19 govorcev spola ni bilo možno nedvoumno določiti.

Posnetki so bili ročno segmentirani in transkribirani. Hkrati je bilo označeno tudi akustično ozadje in negovorni akustično dogodki. To je posledica produkcije oddaj, saj je pogosto v ozadje zvočnega posnetka glavnega govorca montiran zvočni posnetek iz videa ali pa drugo zvočno ozadje, kot je na primer glasba. Pri avtomatskem razpoznavanju govora je pomemben vidik tudi, ali gre za bran, načrtovan ali spontan govor, saj ta značilnost pomembno vpliva na dosežene rezultate.

V predhodnem odstavku naštete parametre v domeni razpoznavanja govora televizijskih oddaj karakterizirajo F-razredi (Schwartz et al., 1997). Ti so definirani na sledeči način:

- F0: bran govor v studijskem okolju,
- F1: spontan govor v studijskem okolju,
- F2: bran/spontan govor preko telefona,
- F3: bran/spontan govor z glasbo v ozadju,
- F4: bran/spontan govor z drugim zvočnim ozadjem,
- F5: govorci, katerih materni jezik ni slovenščina,
- FX: preostalo.

Predstavljene F-razrede bomo uporabili pri podrobnejši analizi rezultatov v šestem poglavju, saj bodo služili za oceno težavnosti testnega scenarija. F-razredi so v govorni bazi zastopani v različnih deležih. Ker predstavlja testni nabor v dolžini 3 ur manj kot eno desetino baze, se to odraža tudi v zastopanosti F-razredov. Tako v testni množici v celoti manjka razred F5 z govorci katerih materni jezik ni slovenščina. Po obsegu pa je najmanjši razred F2 s telefonskim pogovorom. Vsebuje samo 8 segmentov 3 govorcev, ki skupaj izgovorijo nekaj več kot 100 besed.

Nabor učne množice za akustično modeliranje avtomatskega razpoznavalnika govora smo razširili z interno bazo IETK-TV. Ta baza predstavlja nadgradnjo baze UMB BNSI Broadcast News in je nastala na osnovi istih specifikacij. Obsega 29 ur transkribiranih posnetkov 784 govorcev, ki so v celoti namenjeni akustičnemu modeliranju. Nabor različnih televizijskih oddaje je v bazi IETK-TV razširjen, saj so vključeni tudi intervjuji in okrogle mize. Posledično je delež spontanega govora v bazi IETK-TV več kot enkrat večji kot v bazi UMB BNSI Broadcast News.

Za gradnjo novega jezikovnega modela učnega korpusa nismo razširjali. Uporabili smo obstoječe korpuse BNSI-Speech (573k besed), BNSI-Text (11M besed) in FidaPLUS (621M besed) (Arhar in Gorjanc, 2007). Korpus Večer smo iz učenja izločili, saj so njegovi članki vsebovani v korpusu FidaPLUS.

5. Eksperimentalni sistem

5.1. Akustično modeliranje

Za izgradnjo avtomatskega razpoznavalnika govora smo uporabili odprtokodno orodje Kaldi (Povey et al., 2011), ki omogoča izgradnjo sistema s pristopi globokega učenja.

Izvorni signal, ki je del uporabljene govorne baze, smo najprej oknili, nato pa tvorili značilke v obliki mel-frekvenčnih kepstralnih koeficientov (MFCC). Posamezni vektor značilk je imel 13 elementov, ki smo jim dodali še prvi in drugi odvod. Sledil je postopek akustičnega modeliranja. V primeru orodja Kaldi gre za hibridni pristop, ki v prvem koraku uči prikrite modele Markova, v drugem koraku pa globoko nevronsko mrežo. Kot osnovno enoto za akustično modeliranje smo ponovno uporabili slovenske grafeme, saj smo tako dosegli primerljivost s predhodno objavljenimi rezultati.

Prikriti modeli Markova, uporabljeni v akustičnem modeliranju, imajo tristanjsko levo-desno topologijo. Izgradnja akustičnih modelov poteka postopoma, kjer se koraki učenja parametrov modela z Baum-Welchevo re-estimacijo izmenjujejo s koraki prisilne poravnave učnih transkripcij. Za monofonske akustične modele smo uporabili 40 iteracij, za kontekstno odvisne trifonske modele pa 35 iteracij.

Sledilo je učenje globokih nevronskih mrež. Pri tem smo kot arhitekturo uporabili navadno usmerjeno globoko nevronsko mrežo s p-norm aktivacijsko funkcijo (Zhang et al., 2014). Vrednost parametra *p* smo nastavili na 2, saj smo tako dobili primerjalno najboljše rezultate. Podoben eksperiment smo ponovili v povezavi s spreminjanjem konfiguracije nevronske mreže, kjer se je pokazalo, da je smiselno uporabiti tri skrite plasti. Učenje nevronske mreže je potekalo v 10 regularnih epohah in 5 dodatnih, kar je skupaj predstavljajo 345 učnih iteracij. Predstavljeni parametri so v veliki meri odvisni tako od količine učnega materiala kot tudi od njegove raznolikosti. Posledično jih je potrebno ustrezno prilagoditi za vsak govorni vir. Cilj je doseči dobre rezultate razpoznavanja govora, hkrati pa ohraniti zmožnost generalizacije na nove testne vzorce. V nasprotnem primeru dosežemo prekomerno prileganje globoke nevronske mreže. Omejena količina razpoložljivega učnega govornega materiala je bila tudi razlog, da nismo uporabili kompleksnejših pristopov globokega učenja, kot so na primer »end-to-end« globoke nevronske mreže.

5.2. Jezikovno modeliranje

V eksperimentih smo uporabili 2 slovarja, prvi je vseboval 64.000 besed, drugi pa 250.000. Pripadajoča slovarja izgovorjav smo tvorili na osnovi grafemskih akustičnih enot, katerim smo dodali model tišine in pa ločen model različnih negovornih zvokov, ki jih je tvoril govorec. Prvi slovar je identičen slovarju iz prejšnjih eksperimentov (Žgank in Sepesy Maučec, 2010; Žgank et al., 2014), drugi pa vsebuje vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 250.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa FidaPLUS. Delež besed izven slovarja (OOV) prvega slovarja je 4,22 %, drugega pa 1,33 %. Ker oba slovarja vsebujeta besede iz korpusa BNSI-Speech, so med običajnimi besedami tudi različna mašila in onomatopeje, ki smo jih modelirali kar na osnovi njihove zvočne pojave in ne kot posebne, ločene, akustične modele.

Z orodjem SRI Language Modeling Toolkit (Stolcke, 2002) smo zgradili trigramske modele. Trigramski model s prvim slovarjem je identičen kot v prejšnjih eksperimentih (Žgank in Sepesy Maučec, 2010, Žgank et al., 2014). Tudi z drugim slovarjem smo zgradili interpoliran trigramski model. V vseh treh komponentah smo uporabili Good-

Turingovo glajenje in sestopanje po Katz-u. V komponenti BNSI-text smo izločili trigrame s frekvenco 1, v komponenti FidaPLUS pa bigrame s frekvenco 1 in trigrame s frekvencama 1 in 2. Na ta način smo dobili trigramski model, ki je bil primerljive velikosti kot trigramski model s prvim slovarjem. Perpleksnost modela na testni množici je bila 284.

6. Rezultati razpoznavanja govora

Vrednotenje različnih sistemov avtomatskega razpoznavanja govora smo izvedli na testni množici baze UMB BNSI Broadcast News (BNSI-eval), ki vsebuje 4 televizijske oddaje v obsegu 3 ur. Za metriko vrednotenja uspešnosti razpoznavanja govora smo uporabil delež napačno razpoznanih besed (Word Error Rate – WER), ki je definiran kot:

$$WER(\%) = \frac{(I+D+S)}{N} \cdot 100\,, \qquad (1)$$

kjer je *I* število vrinjenih besed, *D* število izbrisanih besed in *S* število zamenjanih besed. *N* predstavlja število vseh besed v testni množici. V delu analize rezultatov smo uporabili kot metriko tudi delež napačno razpoznanih lem (Lemma Error Rate – LER), ki je definiran kot:

$$LER(\%) = \frac{(i+d+s)}{n} \cdot 100\,, \qquad (2)$$

kjer je *i* število vrinjenih lem, *d* število izbrisanih lem in *s* število zamenjanih lem. *n* je skupno število vseh lem v testni množici in je enako številu besed *N*.

V prvem koraku evalvacije smo izvedli primerjavo med avtomatskim razpoznavalnikom govora s prikritimi modeli Markova in globokimi nevronskimi mrežami. Pri tem je sistem s prikritimi modeli Markova služil za primerjavo z rezultati sistema iz leta 2014, ki je takrat dosegel najboljši WER 26,81 % (Žgank et al., 2014). Rezultati napake razpoznavanja besed s trigramskim jezikovnim modelom in slovarjem besed z velikostjo 64k so predstavljeni v tabeli 1.

Sistem	WER [%]
HMM	26,42
DNN	20,85

Tabela 1: Rezultati razpoznavanja govora s trigramskim 64k jezikovnim modelom.

Izhodišča primerjava HMM sistema med letoma 2014 in 2020 kaže, da je prehod na novo ogrodje za avtomatsko razpoznavanje govora potekal brez težav, saj smo dosegli zelo primerljiv WER (iz 26,81 % na 26,42 %). Trenutne HMM akustične modele je sicer možno dodatno nadgraditi s pristopoma od govorca neodvisne transformacije značilk z uporabo LDA (angl. Linear Discriminant Analysis) in MLLT (angl. Maximum Likelihood Linear Transform) (Gales, 1999), kar izboljša rezultat na 24,48 %. Vendar je to izboljšanje relativno omejeno v primerjavi z možnostmi, ki jih v ustreznih pogojih omogoča globoko učenje. Prehod na globoke nevronske mreže za akustično modeliranje izboljša napako razpoznavanja besed na 20,85 %, kar predstavlja pomembno razliko. Pri tem je potrebno posebej izpostaviti, da je količina govornega učnega materiala relativno omejena za pristope globokega učenja.

V drugem koraku smo izvedli vrednotenje, kako vpliva na rezultate izboljšani jezikovni model z bistveno večjim slovarjem besed. Prehod iz 64k besed na 250k besed namreč izdatno zniža delež besed izven slovarja, in ga približa deležu, ki ga najdemo v tipičnih jezikovnih modelih za angleški jezik pri velikosti slovarja 64k. Se pa poveča perpleksnost takšnega jezikovnega modela. Rezultati razpoznavanja govora z akustičnimi modeli DNN in obema trigramskima jezikovnima modeloma so predstavljeni v tabeli 2.

Jezikovni model	WER [%]
64k-3g	20,85
250k-3g	17,19

Tabela 2: Rezultati razpoznavanja govora z akustičnimi modeli DNN in z različnima trigramskima jezikovnima modeloma.

Tudi v tem scenariju razpoznavanja govora je prišlo do znatnega zmanjšanja napake razpoznavanja besed, saj je WER znašal 17,19 %. S povečanjem slovarja razpoznavalnika govora smo tako izboljšali delovanje za 3,66 %, kar je primerljivo z zmanjšanjem deleža OOV. Pri tem smo ohranili kompleksnost sistema na primerljivi ravni, za kar smo poskrbeli med procesom izdelave jezikovnega modela. Razpoznavanje slovenščine z nevronskimi mrežami so predstavili tudi Ulčar et al., 2019. Dosegli so WER 27,16 % na bazi GOS VideoLectures 2.0. Zaradi uporabe različnih govornih in jezikovnih virov rezultati niso medsebojno primerljivi.

V nadaljevanju poglavja bomo podrobneje predstavili analizo doseženih rezultatov razpoznavanja govora. Odgovoriti poskušamo na vprašanje, kako različni faktorji vplivajo na WER. V to skupino sodijo delež besed izven slovarja, pregibna oblika besed, akustični ozadje in način govora.

Referenčne transkripcije in rezultate razpoznavanja smo oblikoslovno označili ter lematizirali z označevalnikom slovenskega jezika Obeliks (Grčar et al., 2012). Oznake besedne vrste in leme so nam koristile pri podrobnejši analizi rezultatov.

S primerjavo lematizirane referenčne transkripcije ter lematiziranih rezultatov razpoznavanja govora smo določili delež napačno razpoznanih lem ter izluščili napake, kjer je lema pravilno razpoznana, besedna oblika pa ne. Na takšen način smo lahko delno analizirali vpliv pregibnosti slovenskega jezika na rezultate razpoznavanja govora. S pomočjo oblikoslovnih oznak pa smo nato še napake v besedni obliki razdelili po besednih vrstah.

V tabeli 3 so predstavljeni podrobnejši rezultati. Razdeljeni rezultati po F-razredih in po spolu kažejo večinoma podobna izboljšanja pri prehodih med sistemi. Opazna je razlika med rezultati za moške in ženske govorce, ki znaša 4,56 %. To razliko bo potrebno še podrobneje analizirati v prihodnosti. Večja izboljšanja vidimo v razredih F1, F3 in FX pri prehodu iz sistema HMM na DNN ter pri razredu F2 pri prehodu na večji slovar, ki pa predstavlja le zelo majhen del testne množice. Med tem ko za bran studijski govor dosegamo WER 9,14 %, sprememba na spontani govor ali dodajanje akustičnega ozadja poslabša rezultate v rangu 10 do 15 %. Pri tem je pričakovano poslabšanje večje, v kolikor je v ozadju dodana glasba.

Sistem	HMM 64k-3g	DNN 64k-3g	DNN 250k-3g
WER [%]	26,42	20,85	17,19
WER - F0 [%]	14,89	12,35	9,14
WER - F1 [%]	35,37	26,70	23,72
WER - F2 [%]	53,38	51,69	33,89
WER - F3 [%]	38,53	28,60	24,13
WER - F4 [%]	29,70	23,71	19,52
WER - FX [%]	34,84	26,54	23,83
WER - Moški [%]	28,92	22,57	19,21
WER - Ženske [%]	23,26	18,57	14,65
LER [%]	23,66	18,07	14,85
WER – LER	2,76	2,78	2,34

Tabela 3: Podrobnejša predstavitev rezultatov razpoznavanja po F-razredih in spolu ter rezultati pravilnosti razpoznave lem.

Rezultati deleža napačno razpoznanih lem LER so po pričakovanjih nižji od rezultatov WER. Te razlike nakazujejo napake v razpoznavanju, kjer je sistem napačno razpoznal besedno obliko, vendar imata tako razpoznana kot pravilna beseda enako lemo. Vidimo, da je razlika manjša pri sistemu z večjim slovarjem, kar nakazuje, da je za del napačno razpoznanih besednih oblik odgovoren omejen slovar.

Treba je dodati, da je v nekaterih primerih bila razpoznana pravilna besedna oblika, vendar je lematizator označil različni lemi med hipotezo in referenco. Ti primeri so se šteli kot napake v vrednotenju LER. To so dogaja predvsem pri primerih, kjer se zaradi drugih napak (izbrisanih ali vrinjenih kratkih besed) spremeni kontekst besede. Na primer, besedna oblika *ukrepa* je lahko označena z lemo *ukrep* (samostalnik) ali pa *ukrepati* (glagol). Ocenjujemo pa, da je delež teh primerov le majhen. Iz tega sklepamo, da je delež napak, ki so posledica pregibnosti jezika, nekoliko višji kot pa razlika med WER in LER, namreč med 2,5 in 3 %.

V nadaljevanju smo pregledali napake v besedni obliki pri isti lemi glede na besedno vrsto. Rezultati so podani v tabeli 4. Podali smo le pregibne besedne vrste (brez zaimkov). Primerjamo sistem HMM 64k-3g in DNN 250k-3g. Vidimo, da je le relativno izboljšanje pri napačno razpoznanih oblikah številnikov primerljivo z relativnim izboljšanjem skupnega rezultata, ki je 34,9 %. Najmanjše relativno izboljšanje pa vidimo pri glagolih. Skupno relativno izboljšanje napak v besedni obliki pa je približno dvakrat manjše od relativnega izboljšanja skupnega rezultata.

Rezultati kažejo na to, da sistem s povečanim slovarjem in uporabo nevronskih mrež pomembno zmanjša skupni delež napak razpoznavanja. Vidimo pa, da je relativno zmanjšanje napak zaradi pregibnosti besed manjše glede na skupno zmanjšanje. V sistemu DNN 250k-3g je tako delež napak zaradi pregibnosti 13,3 %, kar je več kot pri sistemu HMM, kjer je ta delež 10,5 %.

Pregled posameznih najpogostejših parov zamenjav ne kaže zanimivih rezultatov glede pregibnih besed. Večinoma se v pogostih parih zamenjav pojavljajo kratke besede (npr. zamenjave so – se, na – no ipd.). Najpogostejši par zamenjave, kjer je prišlo do napake v besedni obliki polnopomenske pregibne besede je par stališče–stališča, ki se pojavi štirikrat v sistemu DNN 250k-3g.

7. Zaključek

V članku smo predstavili sistem za avtomatsko razpoznavanje govora v domeni televizijskih oddaj, ki je dosegel najboljši delež napake razpoznavanja besed 17,19 %. Izboljšanje je v pretežni meri rezultat uporabe akustičnih modelov z globokimi nevronskimi mrežami in vpliva zmanjšanja deleža besed izven slovarja. Z večanjem slovarja smo uspešno zmanjšali vpliv pregibnosti slovenskega jezika.

Podrobnejša analiza po F-razredih in lemah je pokazala, da je nadaljnje izboljšanje rezultatov možno doseči predvsem na račun izboljšanja akustičnega modeliranja v primeru kratkih besed in govora v zahtevnejših pogojih.

Zahvala

Zahvaljujemo se avtorjem besedilnega korpusa FidaPLUS, ki so nam omogočili njegovo uporabo za jezikovno modeliranje avtomatskega razpoznavalnika govora.

Besedna vrsta	Št. napak v HMM 64k-3g	Št. napak v DNN 250k-3g	Relativna izboljšava [%]
Samostalnik	309	265	14,2
Pridevnik	155	112	27,7
Glagol	148	135	8,8
Števnik	16	10	37,5
Prislov	0	1	-
SKUPAJ	628	522	16,9

Tabela 4: Napačno razpoznane besedne oblike glede na besedno vrsto.

8. Literatura

Špela Arhar in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2., 95—110.

Simon Dobrišek, Jerneja Gros, France Mihelič in Nikola Pavešić. 1998. Recording and labelling of the GOPOLIS Slovenian speech database. V: *First International Conference on language resources & evaluation*: Granada, Spain, 28-30 May 1998 (str. 1089—1096). European Language Resources Association.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič in Nikola Pavešić. 2017. Speech Database of Spoken Flight Information Enquiries SOFES 1.0.

ELRA. 2020. BNSI Catalog Reference : S0275: www.elra.info.

Mark J. Gales. 1999. Semi-tied covariance matrices for hidden Markov models. *IEEE transactions on speech and audio processing*, 7(3), 272—281.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenija.

Janez Kaiser in Zdravko Kačič. 1997. SpeechDat (II) Slovenian Database for the Fixed Telephone Network. Maribor, Slovenia: University of Maribor.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel … in Jan Silovsky. 2011. The Kaldi speech recognition toolkit. V: *IEEE ASRU 2011 Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Richard Schwartz, Hubert Jin, Francis Kubala in Spyros Matsoukas. 1997. Modeling those F-Conditions - or not. *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. *International Conference on Speech and Language Processing*, II: 901—904.

Matej Ulčar, Simon Dobrišek in Marko Robnik-Šikonja. 2019. Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika. 27, 3*.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation*, 47(4), 1031—1048.

Darinka Verdonik. 2018. Korpus in baza Gos Videolectures. *Zbornik 11. konference Jezikovne tehnologije*, Informacijska družba-IS, Ljubljana.

Xiaohui Zhang, Jan Trmal, Daniel Povey in Sanjeev Khudanpur. 2014. Improving deep neural network acoustic models using generalized maxout networks. V: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (str. 215—219). IEEE.

Andrej Žgank, Zdravko Kačič in Bogomir Horvat. 2002. Preliminary evaluation of Slovenian mobile database PoliDat. V: *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'02).

Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, Darinka Verdonik, Jani Kitak, Damjan Vlaj, Vladimir Hozjan, Zdravko Kačič in Bogomir Horvat. 2004. Acquisition and annotation of Slovenian broadcast news database. *Fourth international conference on language resources and evaluation*, LREC 2004, Lizbona, Portugalska.

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj, in Zdravko Kačič. 2006. Sloparl-Slovenian parliamentary speech and text corpus for large vocabulary continuous speech recognition. V: *Ninth International Conference on Spoken Language Processing.*

Andrej Žgank in Mirjam Sepesy Maučec. 2010. Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Jezikovne tehnologije 2010*, Ljubljana, Slovenija.

Andrej Žgank, Gregor Donaj in Mirjam Sepesy Maučec. 2014. Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov. *Zbornik 9. konference Jezikovne tehnologije, Informacijska družba-IS* (str. 147—150).

Andrej Žgank, Mirjam Sepesy Maučec in Darinka Verdonik. 2016. The SI TEDx-UM speech database: A new Slovenian spoken language resource. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (str. 4670—4673).

Janez Žibert, Sanda Martinčić-Ipšić, Ivo Ipšić in France Mihelič. 2003. Bilingual speech recognition of Slovenian and Croatian weather forecasts. V: Proceedings EC-VIP-MC 2003. *4th EURASIP Conference focused on Video/Image Processing and*

Multimedia Communications (IEEE Cat. No. 03EX667) (Vol. 2, str. 637—642). IEEE.

Janez Žibert in France Mihelič. 2004. Development of Slovenian broadcast news speech database. V: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).*