AVTOMATSKO RAZPOZNAVANJE SLOVENSKEGA GOVORA ZA DNEVNOINFORMATIVNE ODDAJE

Lucija GRIL, Mirjam SEPESY MAUČEC, Gregor DONAJ, Andrej ŽGANK
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Gril, L., Sepesy Maučec, M., Donaj, G., Žgank, A. (2021): Avtomatsko razpoznavanje slovenskega govora za dnevnoinformativne oddaje. Slovenščina 2.0, 9(1): 60–89.

DOI: https://doi.org/10.4312/slo2.0.2021.1.60-89

Na področju govornih in jezikovnih tehnologij predstavlja avtomatsko razpoznavanje govora enega izmed ključnih gradnikov. V prispevku bomo predstavili razvoj avtomatskega razpoznavalnika slovenskega govora za domeno dnevnoinformativnih oddaj. Arhitektura sistema je zasnovana na globokih nevronskih mrežah. Pri tem smo ob upoštevanju razpoložljivih govornih virov izvedli modeliranje z različnimi aktivacijskimi funkcijami. V postopku razvoja razpoznavalnika govora smo preverili tudi, kakšen je vpliv izgubnih govornih kodekov na rezultate razpoznavanja govora. 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 250.000 besed. Na ta način smo znižali delež besed izven slovarja na 1,33 %. Z razpoznavanjem govora na testni množici smo dosegli najboljšo stopnjo napačno razpoznanih besed (WER) 15,17 %. Med procesom vrednotenja rezultatov smo izvedli tudi podrobnejšo analizo napak razpoznavanja govora na osnovi lem in F-razredov, ki v določeni meri pokažejo na zahtevnost slovenskega jezika za takšne scenarije uporabe tehnologije.

Ključne besede: avtomatsko razpoznavanje slovenskega govora, lastnosti slovenskega jezika, dnevnoinformativne oddaje, globoke nevronske mreže, izgubni govorni kodeki.

1 UVOD

V zadnjem desetletju spremljamo izredno hiter razvoj področja umetne inteligence, ki mu botruje predvsem tehnološki napredek na področju velikih podatkov in algoritmov za globoko učenje. To je pripeljalo tudi do izboljšanja metod na področju govornih in jezikovnih tehnologij. Strateški cilji države se lahko tako učinkovito osredotočajo na vključujočo družbo, ki uspešno uporablja tehnologije digitalizacije. Naravna interakcija med človekom in napravami v inteligentnem okolju je eden izmed ključnih vidikov sprejemljivosti tehnologije.

Splošna razširjenost pametnih naprav, kot so mobilni telefoni, je prispevala k povečanju količin različnega zvočnega (in slikovnega) gradiva, ki je na voljo uporabniku. V želji zagotoviti učinkovit dostop do informacij, ki jih vsebuje takšna množica zvočnega gradiva, je neobhodno potrebna uporaba tehnoloških rešitev.

Ena izmed jedrnih tehnologij, ki omogočajo ustrezno podporo za zajemanje informacij, tako iz uporabniškega ali medijskega zvočnega toka kot tudi iz uporabniškega vmesnika naprav v inteligentnem okolju, 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. Na področju avtomatskega razpoznavanja govora je do pomembnih korakov v razvoju prišlo na točki, ko je bilo možno za to nalogo učinkovito uporabiti globoke nevronske mreže. Te so zamenjale prejšnjo arhitekturo, ki je temeljila na prikritih modelih Markova in zadnja leta ni več prinašala bistvenega napredka. Metode globokega učenja danes predstavljajo privzeto arhitekturo na praktično vseh področjih govornih in jezikovnih tehnologij.

Pomemben vidik predstavlja tudi računska zahtevnost, ki lahko pogosto trči ob vprašanja zagotavljanja zasebnosti govorcev, kadar je v uporabi procesiranje v oblaku. Ta vidik je lahko izrednega pomena, kadar govorimo o tehnologijah za vključujočo družbo, ki pogosto pokrivajo zelo osebne vidike komunikacije uporabnikov.

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. To se lahko še dodatno potencira s posebnostmi določenih jezikov, ki otežijo avtomatsko razpoznavanje govora. V kategorijo za procesiranje zahtevnih jezikov sodi tudi slovenščina. Zanjo je značilna visoka pregibnost besed in relativno prost vrstni red besed v stavku. Obe lastnosti pomembno vplivata na rezultate razpoznavanja govora, saj prvič povečata akustično zamenljivost besed in iskalni prostor razpoznavalnika, drugič pa zmanjšata predikcijsko zmožnost statističnih jezikovnih modelov.

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 intenzivnemu razvoju v svetu. 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 Razvoj slovenščine v digitalnem okolju (RSDO, b.d.), ki bo potekal do leta 2023.

Cilj pričujočega raziskovalnega dela je predstaviti razvoj sistema za avtomatsko razpoznavanje slovenskega govora z globokimi nevronskimi mrežami, ki deluje za domeno dnevnoinformativnih oddaj. Takšen avtomatski razpoznavalnik govora je lahko zelo pomembno govornotehnološko orodje za različne scenarije uporabe, kot so na primer avtomatsko indeksiranje govorne vsebine, avtomatsko podnaslavljanje ali avtomatsko prevajanje govora v govor. Za učinkovito doseganje teh ciljev je treba uporabljati razpoznavalnike govora z metodami globokega učenja. Dosedanji sistemi za avtomatsko razpoznavanje slovenskega govora za domeno dnevnoinformativnih oddaj (Žgank in Sepesy Maučec, 2010; Žgank idr., 2014) so temeljili na predhodni arhitekturi prikritih modelov Markova.

V prispevku želimo podati oceno, kakšen primanjkljaj pri prehodu na novo arhitekturo globokih nevronskih mrež predstavljajo omejene govorne baze za slovenski jezik. Pri izgradnji modelov smo se odločili za uporabo različnih aktivacijskih funkcij nevronskih mrež ter na ta način izvedli primerjavo arhitektur. Podoben potek eksperimenta razvoja razpoznavalnika govora so uporabili za španski jezik (Zorrilla idr., 2016), kjer so bile izhodišče obstoječe metode, ki so jih nato preverili na že predhodno uporabljenih govornih

bazah za španski jezik. Hkrati nas v okviru raziskave zanima tudi, kakšen vpliv ima uporaba izgubnih kodekov na rezultate avtomatskega razpoznavanja govora. Izgubni kodeki so postali pomembni že z razmahom različnih internetnih pretočnih storitev. Še posebej velik pomen pa so dobili v času epidemije covida-19, ko se je večina komuniciranja in funkcioniranja družbe preselila v oddaljen način. Podobno primerjavo vpliva izgubnega kodiranja na rezultate razpoznavanja govora sta za drugo domeno in jezik izvedla Pollak in Behunek (2011). V zadnjem delu prispevka bomo izvedli tudi analizo napak razpoznavanja govora in na ta način poskušali ugotoviti vpliv visoke pregibnosti na rezultate razpoznavanja govora. Raziskovalno delo smo zasnovali na slovenski bazi televizijskih dnevnoinformativnih oddaj UMB BNSI Broadcast News (Žgank idr., 2004) in IETK-TV (Žgank idr., 2014), saj ti govorni bazi trenutno še vedno predstavljata najprimernejši vir za takšno analizo, hkrati pa omogočata tudi primerljivost rezultatov s starejšimi sistemi avtomatskega razpoznavanja govora.

V nadaljevanju članka bomo najprej predstavili trenutno stanje na področju razpoznavanja govora za slovenski jezik. V tretjem poglavju bo sledila kratka predstavitev teoretičnega ozadja metod, ki se danes uporabljajo pri gradnji avtomatskih razpoznavalnikov govora. Opisali bomo tudi področje govornih kodekov. V četrtem poglavju bomo predstavili uporabljene govorne in jezikovne vire. 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 PODROČJA AVTOMATSKEGA RAZPOZNAVANJA GOVORA ZA SLOVENSKI JEZIK

2.1 Govorni viri za slovenski jezik

Že v uvodu smo zapisali, da predstavljajo govorni viri ključno komponento za razvoj avtomatskega razpoznavalnika govora. Pomembno je, da 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 idr., 2002), Gopolis (Dobrišek idr., 1998), VNTV/VNRAD (Žibert idr., 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 idr., 2004), SiBN Broadcast News (36 ur) (Žibert in Mihelič, 2004), IETK-TV (30 ur) in GOS javni podkorpus (42 ur) (Verdonik idr., 2013). Predavanja najdemo v bazi SI TEDx-UM (54 ur, avtomatske transkripcije) (Žgank idr., 2016) in bazi GOS-VideoLectures (22 ur) (Verdonik, 2018). Baza SloParl (Žgank idr., 2006) vsebuje 100 ur posnetkov in magnetogramov parlamentarnih razprav iz DZ RS, baza SOFES (Dobrišek idr., 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. Z vidika razvoja področja avtomatskega razpoznavanja govora za slovenski jezik predstavlja takšna razdrobljena dostopnost velik izziv.

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.

2.2 Avtomatsko razpoznavanje govora za slovenski jezik

V nadaljevanju bomo podali še kratek pregled ključnih aktivnosti na področju avtomatskega razpoznavanja slovenskega govora. Raziskave so začele potekati okoli leta 1990. Prvi sistemi razpoznavanja govora so delovali za preprostejše scenarije, kot so: krmiljenje preprostih aplikacij (Kačič idr., 1988), klasifikacija fonemov (Mihelič idr., 1992) ali razpoznavanje števk (Imperl idr., 1996). V naslednjem koraku so sledili zahtevnejši scenariji, ki temeljijo na vezanih besedah – dialog za poizvedovanje o letalskih informacijah (Ipšić idr., 1999) ter poizvedovanje o telefonskih številkah (Imperl in Kačič, 1999). Prehod na scenarije razpoznavanja tekočega govora z velikim slovarjem besed (Kaiser idr., 2000) prvič pokaže na izzive, povezane s kompleksnostjo visokopregibnega slovenskega jezika, ter težave zaradi ne dovolj razvitih govornih virov. Delno je to možno izničiti z omejitvijo na ozko domeno, kot so na primer vremenske napovedi (Žibert idr., 2000). Pomembnejši pa je bil korak v smeri razvoja novih govornih virov s področja dnevnoinformativnih oddaj (Žgank idr., 2004; Žibert in Mihalič, 2004), ki so potem služile za razvoj kompleksnejših razpoznavalnikov tekočega govora (Žgank idr., 2006; Dobrišek in Mihelič, 2010; Žgank in Sepesy Maučec, 2010; Žgank idr., 2014).

Prvi slovenski razpoznavalnik govora z globokimi nevronskimi mrežami je bil razvit v okviru večjezičnega razpoznavanja za južnoslovanske jezike (Nouza idr., 2016). V zadnjem desetletju postaja na področju razpoznavanja govora poleg domene dnevnoinformativnih oddaj pomembna tudi domena predavanj. K temu je v veliki meri pripomogel razvoj multimedijske tehnologije in priljubljenost masovnih spletnih predavanj (MOOC). Tako pride tudi

v slovenskem prostoru do izgradnje ustreznih govornih baz s tega področja (Zwitter Vitez idr., 2013; Verdonik idr., 2017). Avtomatski razpoznavalnik govora z globokimi nevronskimi mrežami, ki deluje za to domeno, je predstavil Ulčar s sodelavci (2019) in vključuje sledeče govorne vire: GOS 1.0 (Zwitter Vitez idr., 2013), Gos VideoLectures 2.0 (Verdonik idr., 2017) in Sofes 1.0 (Dobrišek idr., 2017).

3 ARHITEKTURE ZA AVTOMATSKO RAZPOZNAVANJE 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 predstavljajo metodo statističnega modeliranja, kjer na osnovi vhodnih vektorjev značilk ocenjujemo 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 metodo 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. Kadar je arhitektura nevronske mreže načrtovana tako, da vsebuje dve ali več plasti, govorimo o globoki nevronski mreži. Število globokih plasti, ki jih uporabimo v postopku strojnega učenja, je v veliki meri odvisno od količine učnega gradiva.

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

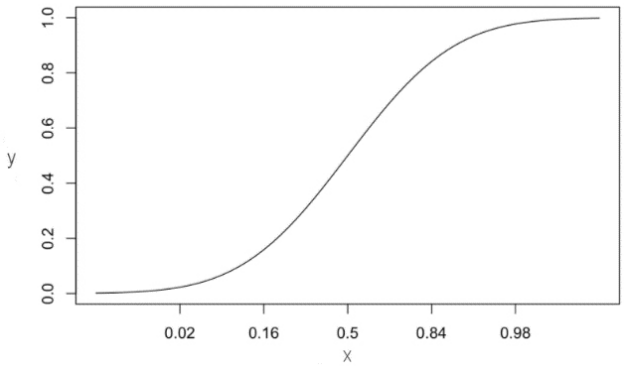
Aktivacijske funkcije so lahko različnih tipov – stopničaste, linearne ali nelinearne. Stopničasta aktivacijska funkcija temelji na pragovni vrednosti (angl. threshold). Če je vhodna vrednost nad ali pod določenim pragom, se nevron aktivira in pošlje naslednji plasti povsem enako vrednost. Linearna funkcija vzame vhodno vrednost nevrona, jo pomnoži z utežjo in generira izhodni signal. Nelinearne aktivacijske funkcije omogočajo kompleksnejše preslikave vhodnih vrednosti v izhodne.

Tanh je hiperbolična tangenta funkcija, ki jo uporabljamo kot aktivacijsko funkcijo pri globokih nevronskih mrežah. Zaloga vrednosti funkcije je med –1 in 1, zaradi česar je povprečje skrite plasti 0 ali blizu te vrednosti. To pomeni, da je učenje na naslednji plasti veliko lažje.

P-norm je nelinearna aktivacijska funkcija, katere izhod se izračuna kot:

$$y = (\sum_i |x_i|^p)^{1/p}, \tag{1}$$

kjer so vektorji x majhna skupina vhodnih vrednosti. Vrednost p je spremenljiva in zanjo je bilo pokazano (Zhang idr., 2014), da s p = 2 pridobimo najboljše rezultate.



Slika 1: Graf aktivacijske funkcije p-norm.

Pri načrtovanju arhitekture globokih nevronskih mrež lahko dodamo ozka podatkovna grla, ki jih bomo v nadaljevanju navajali kar kot ozka grla. Ozko grlo je plast, ki ima manj nevronov kot plast pred ali za njo. Takšne plasti spodbudijo, da se značilke bolje prilagodijo razpoložljivemu prostoru parametrov, ki ga omejimo z velikostjo ozkega grla. Z ozkim grlom tako dosežemo predstavitev vhoda z manjšo dimenzijo.

Prav tako se pri načrtovanju arhitekture uporabljajo razne oblike ansambla. Ideja ansambla je, da namesto enega klasifikatorja zgradimo več klasifikatorjev, ki na koncu glasujejo o končni odločitvi. Učenje poteka na enakih učnih podatkih za vsako iteracijo. Po vsaki iteraciji se doda vrednost, ki je zmnožek vrednosti β in križne entropije izhoda trenutne mreže ter geometrijsko povprečnih zadnjih vrednosti izhoda ansambla mrež. Vrednost β eksponentno narašča glede na začetno in končno vrednost β, ki jo izberemo.

V zadnjih letih so nevronske mreže postale popularne na raznih področjih strojnega učenja, tudi pri razpoznavanju govora (Nassif idr., 2019). Ker pa gre pri razpoznavanju govora za razpoznavanje časovne vrste, vse arhitekture nevronskih mrež niso 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. V nadaljevanju bomo preizkusili, kako dobro delujejo glede na našo učno množico različne nelinearne aktivacijske funkcije, ki so uporabljene pri izgradnji razpoznavalnikov govora. Pogosto uporabljeni sta p-norm in tanh, ki smo ju kombinirali še z ansamblom in ozkim grlom, saj smo želeli preveriti, ali bodo dodatni koraki doprinesli k izboljšanju rezultatov.

3.3 Zvočni kodeki

Za stiskanje podatkov uporabljamo kodiranje, ki nam omogoča, da lahko informacijo zapišemo z manj biti kakor na začetku. Pri zapisu zvoka lahko na takšen način zmanjšamo pasovno širino in velikost stisnjene zvočne datoteke.

Kodiranje je lahko brezizgubno ali izgubno. Brezizgubni zvočni kodeki zmanjšajo obseg podatkov, vendar ohranijo vso informacijo, ki jo lahko ponovno pridobimo po dekodiranju. Pri izgubnih kodekih se odstranjujejo informacije v časovnem in/ali frekvenčnem prostoru, ki jih človek ne more zaznati zaradi psihoakustičnih značilnosti slušne zaznave. Z uporabo izgubnih kodekov se zmanjša bitna ločljivost zvoka, zaradi česar po dekodiranju nikoli ne pridobimo prvotne informacije v celoti. Vpliv popačenj izgubnih kodekov želimo ohraniti na tako nizki ravni, da ne vplivajo bistveno na subjektivno zaznavo kakovosti zvoka.

Izgubni zvočni kodeki so pomembno pridobili na veljavi z razmahom internetnih storitev, še posebej v obliki pretočnega dostopa do vsebin in različnih oblik dela na daljavo v času epidemije covida-19. Posledično moramo upoštevati njihov vpliv tudi na področju avtomatskega razpoznavanja govora.

4 UPORABLJENI GOVORNI IN JEZIKOVNI VIRI

Osrednji vir podatkov, ki smo jih uporabili za akustično modeliranje, je predstavljala govorna baza UMB BNSI Broadcast News (Žgank idr., 2004), ki jo distribuira organizacija ELRA (2015). Govorna baza vsebuje posnetke dnevnoinformativnih 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 akustični 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 idr., 1997). Ti so definirani na sledeč 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. Pomembno namreč odražajo akustično ozadje in s tem nakazujejo na potencialni vpliv degradacij na rezultat razpoznavanja govora. 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, ki vsebuje govor, posnet preko telefona. Ta kategorija vsebuje samo osem segmentov treh govorcev, ki skupaj izgovorijo nekaj več kot 100 besed.

Nabor učne množice za akustično modeliranje avtomatskega razpoznavalnika govora smo razširili še z govorno bazo IETK-TV, ki pa zaradi omejitev avtorskih pravic ni širše dostopna. 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 oddaj je v bazi IETK-TV razširjen v primerjavi z bazo UMB BNSI, 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 jezikovnega modela učnega korpusa nismo razširjali. Uporabili smo sledeče korpuse: BNSI-Speech (573 tisoč besed), BNSI-Text (11 milijonov besed) in FidaPLUS (621 milijonov besed) (Arhar in Gorjanc, 2007). Korpus Večer smo iz učenja izločili, saj so njegovi članki vsebovani v korpusu FidaPLUS.

5 EKSPERIMENTALNI SISTEM

Osnovna zasnova eksperimentalnega sistema za avtomatsko razpoznavanje govora, uporabljena v teh eksperimentih, je enaka za pristopa HMM in DNN. Zajet govorni signal je najprej treba predprocesirati in pretvoriti v vektorje značilk. Nato lahko izvedemo razpoznavanje govora, kjer uporabimo akustične in jezikovne modele ter fonetični slovar. Akustične modele smo s pristopi strojnega učenja predhodno naučili na transkribirani učni govorni bazi, jezikovne modele pa na učnem besedilnem korpusu.

5.1 Akustično modeliranje

Za izgradnjo avtomatskega razpoznavalnika govora smo uporabili odprtokodno orodje Kaldi (Povey idr., 2011), ki omogoča izgradnjo sistema z metodami globokega učenja.

Za začetek učenja akustičnih modelov potrebujemo transkribirane posnetke v formatu WAV. Za učno kot tudi testno množico je treba pripraviti vse spremljajoče datoteke. Za učni postopek smo kot osnovo vzeli Kaldijev postopek učenja z bazo Mini LibriSpeech, ki smo ga ustrezno nadgradili. Uporabljeni postopek učenja je po dosedanjih izkušnjah dajal dobre rezultate, hkrati sta velikosti obeh baz primerljivi.

V naslednjem koraku s pomočjo že pripravljenih skript v orodju Kaldi pripravimo še ostale datoteke, ki so potrebne za učenje akustičnih modelov. Izvorni signal oknimo in nato tvorimo 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, kjer zaporedoma izvajamo učenje modelov in njihove poravnave pred ponovnim učenjem novega modela. V primeru orodja Kaldi gre za hibridno metodo, kjer v prvem koraku uči prikrite modele Markova, v drugem koraku pa globoko nevronsko mrežo. Kot osnovno enoto za akustično modeliranje smo uporabili slovenske grafeme.

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-Welchovo reestimacijo izmenjujejo s koraki prisilne poravnave izboljšanih različic učnih transkripcij. Za monofonske akustične modele smo uporabili 40 iteracij, za kontekstno odvisne trifonske modele pa 35 iteracij učenja.

Sledilo je učenje globokih nevronskih mrež. Pri tem smo kot arhitekturo uporabili navadno usmerjeno globoko nevronsko mrežo. V okviru akustičnega modeliranja smo uporabili različne aktivacijske funkcije. Tako smo preverili možen vpliv arhitekture nevronskih mrež na avtomatsko razpoznavanje slovenskega govora. Prva je bila aktivacijska funkcija p-norm (Zhang idr., 2014). Vrednost parametra p smo nastavili na 2, saj je bilo v preteklosti pokazano (Zhang idr., 2014), da lahko pri tej vrednosti pričakujemo najboljše rezultate.

Učenje nevronske mreže je potekalo v 15 regularnih epohah in 5 dodatnih, kar sta prevzeta parametra za takšen potek. Inicialno stopnjo učenja smo nastavili na 0,02 in končno stopnjo učenja na 0,004. Vhodno število nevronov smo nastavili na 2000 in izhodno število na 400. Nastavljene vrednosti parametrov ustrezajo predlaganim v okolju Kaldi za razpoložljivo količino učnega materiala. V eksperimentu smo implementirali 2 skriti plasti in 4 skrite plasti, saj smo na takšen način prilagajali arhitekturo glede na velikost učnega seta.

Aktivacijsko funkcijo p-norm smo v naslednjem poskusu združili z uporabo ozkega grla. Pri tej kombinaciji s pomočjo nelinearnih vrednosti ustvarjamo značilke ozkega grla. Dimenzijo ozkega grla smo nastavili na 42. Vrednost *p* smo ohranili na 2. Prav tako smo ohranili število epoh in stopnji učenja. Implementirali smo 4 skrite plasti, saj se je iz predhodnega preizkusa izkazalo, da se je model z dvema skritima plastema slabše izkazal. Preizkusili pa smo tudi arhitekturo z nekoliko manj nevroni, in sicer smo vhodno število nevronov nastavili na 1000 in izhodno število nevronov na 200. Tudi v tem primeru smo uporabili 4 skrite plasti.

P-norm smo kombinirali tudi z metodo ansambla. Parametre za p-norm smo za število epoh in vrednost p smo nastavili enako kot v prejšnjih dveh primerih. Prav tako smo tudi tukaj uporabili arhitekturo s 4 skritimi plastmi. Število vhodnih in izhodnih nevronov smo prilagajali tako kot v prejšnjem primeru. V prvem primeru smo uporabili 1000 vhodnih in 200 izhodnih, v drugem pa 2000 vhodnih in 400 izhodnih. Dodali smo parameter velikosti ansambla, ki smo ga nastavili na 4, ter inicialno in končno vrednost β. Inicialno vrednost β smo nastavili na 0,1, končno pa na 5. Te vrednosti so bile nastavljene glede na izhodiščne parametre v okolju Kaldi.

V naslednjem poskusu smo uporabili aktivacijsko funkcijo tanh. Pri tej arhitekturi smo uporabili 20 regularnih epoh in 5 dodatnih. Arhitektura vsebuje dve skriti plasti s 375 nevroni. Tukaj smo enako kot pri aktivacijski funkciji p-norm nastavili inicialno stopnjo učenja na 0,02 in končno stopnjo učenja na 0,004. Tako smo sledili primerljivosti arhitektur.

V kombinaciji aktivacijske funkcije tanh in ozkega grla smo se odločili, da uporabimo enake parametre kot pri globoki nevronski mreži z aktivacijsko funkcijo tanh. Dimenzijo ozkega grla smo nastavili na 42.

Predstavljeni parametri so v veliki meri odvisni tako od količine učnega materiala kot tudi od njegove raznolikosti. Posledično jih je treba ustrezno prilagoditi za vsak govorni vir. Parametre, ki jih nismo vključili v primerjavo, smo nastavili empirično oziroma s pomočjo informacij o sistemih drugih avtorjev. Cilj je doseči dobre rezultate razpoznavanja govora, hkrati pa ohraniti zmožnost posplošitve na nove testne vzorce. V nasprotnem primeru dosežemo prekomerno prileganje globoke nevronske mreže. V takšnem primeru sicer lahko dosežemo izvrsten rezultat razpoznavanja govora na zelo sorodnem testnem gradivu. Kakor hitro pa je testno gradivo raznolikejše, pride do drastičnega poslabšanja rezultatov razpoznavanja govora. Zato je takšno prekomerno prileganje učinek, ki se mu želimo izogniti. Omejena količina razpoložljivega učnega govornega materiala je bila tudi razlog, da nismo uporabili kompleksnejših metod globokega učenja, kot so na primer »end-to-end« globoke nevronske mreže.

5.2 Jezikovno modeliranje

V eksperimentih smo uporabili dva slovarja, prvi je vseboval 64.000 besed, drugi pa 250.000. Pripadajoča slovarja izgovorjav smo tvorili na osnovi grafemskih akustičnih enot, ki smo jim dodali model tišine in pa ločen model različnih negovornih zvokov, ki jih je tvoril govorec. Prvi slovar, ki smo ga naredili z enakim postopkom kot avtorji v prispevkih Žgank in Sepesy Maučec (2010) ter Žgank idr. (2014), obsega 64.000 besed. Vsebuje vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 64.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa Večer. Drugi slovar izhaja iz prvega. Do velikosti 250.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa FidaPLUS. Korpus FidaPLUS smo za razširitev slovarja uporabili zato, ker je to obsežen in reprezentativen korpus splošnega slovenskega jezika. Z razširitvijo slovarja smo želeli zmanjšati delež besed izven slovarja (OOV), ki je v primeru prvega slovarja znašal 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 s prvim slovarjem. Uporabili smo enak potek kot avtorji

v Žgank in Sepesy Maučec (2010) ter Žgank idr. (2014). Tudi z drugim slovarjem smo zgradili interpoliran trigramski model. V vseh treh komponentah smo uporabili Good-Turingovo glajenje in sestopanje po Katzu. 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.

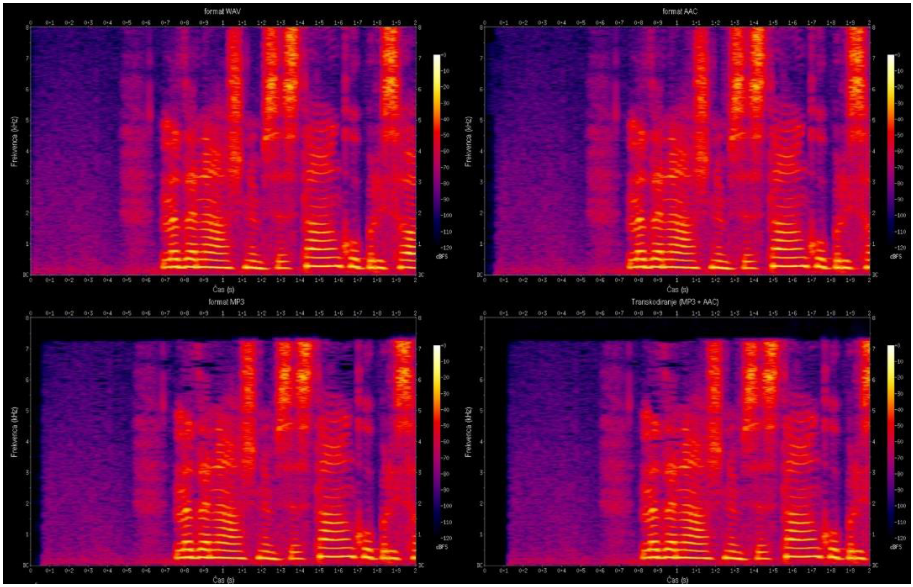
5.3 Izgubno stiskanje govora

Datoteke govorne baze UMB BNSI Broadcast News so v formatu WAV, ki ne uporablja stiskanja zvoka. Zanima nas, kakšno vlogo imajo izgubni kodeki pri avtomatskem razpoznavanju govora. V ta namen smo pripravili nove testne sete zvočnih datotek, ki smo jih najprej pretvorili v format z izgubnim kodekom in potem nazaj v izvorni format, potreben za razpoznavanje govora. V tem delu eksperimenta smo uporabili izgubna kodeka MPEG-1 Audio Layer III (MP3) in njegovega naslednika Advanced Audio Coded (AAC), ki je del skupine kodekov MPEG-2 Part 7. Kodek MP3 je definiran v standardih ISO/IEC 11172-3:1993 in ISO/IEC 13818-3:1995, kodek AAC pa v standardu ISO/IEC 13818-7:1997. Ključna razlika med njima je, da AAC omogoča še bolj učinkovito izgubno stiskanje zvoka pri enakem nivoju človeku zaznavnih degradacij.

Z orodjem SoX smo pretvorili izvorno testno množico datotek iz formata WAV v AAC pri bitni hitrosti 64 kbit/s in 128 kbit/s. Bitna hitrost originalnih datotek v formatu WAV je bila 256 kbit/s. Nato smo ponovili postopek še v obratni smeri in nove stisnjene datoteke pretvorili nazaj v format WAV.

Z orodjem FFmpeg smo pretvorili izvorno testno množico iz formata WAV v MP3 z bitno hitrostjo 64 kbit/s in 128 kbit/s. Postopek smo ponovili še v obratni smeri, da smo iz MP3 pretvorili posnetke nazaj v format WAV.

V naslednjem koraku smo želeli preveriti še, kakšen je vpliv transkodiranja na avtomatsko razpoznavanje govora. V tem primeru gre za večkratno zaporedno kodiranje z izgubnimi kodeki. Vzeli smo testne posnetke v formatu WAV, ki so že bili pretvorjeni v format MP3 z bitno hitrostjo 128 kbit/s, in jih ponovno pretvorili v format AAC z bitno hitrostjo 128 kbit/s in nazaj v format WAV.



Slika 2: Primerjava spektrogramov zvočnega zapisa dolžine 2 sekund v različnih zvočnih formatih.

Na Sliki 2 lahko opazimo, da pride pri formatu MP3 do rezanja frekvenc, višjih od 7,5 kHz, kar je značilno za pretvarjanje v format MP3 pri nizkih bitnih hitrostih. Glede na spektrogram, ki ga dobimo z zvočnim posnetkom formata WAV, lahko na ostalih treh spektrogramih opazimo razlike v deležih spektralne energije v različnih pasovih. Te razlike so nekoliko bolj vidne pri formatu MP3 kakor pri formatu AAC.

Za analizo vpliva izgubnih kodekov na avtomatsko razpoznavanje govora smo uporabili jezikovni model velikosti 64.000 in globoke nevronske akustične modele z aktivacijsko funkcijo tanh, ki so dosegli najboljše rezultate pri testiranju brez izgubne kompresije.

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 uporabili delež napačno razpoznanih besed (Word Error Rate – WER), ki je definiran kot:

$$WER(\%) = \frac{(I+D+S)}{N} \cdot 100, \quad (2)$$

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 kot metriko uporabili tudi delež napačno razpoznanih lem (Lemma Error Rate – LER), ki je definiran kot:

$$LER(\%) = \frac{(i+d+s)}{n} \cdot 100, \quad (3)$$

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 primerjali, kako je spreminjanje parametrov modelov vplivalo v koraku učenja z nevronsko mrežo, ko smo uporabili aktivacijsko funkcijo p-norm. V Preglednici 1 lahko vidimo rezultate WER, ki smo jih dosegli pri razpoznavanju testnega nabora.

Preglednica 1: *Primerjava rezultatov WER glede na različne nastavitve parametrov*

Aktivacijska funkcija	Število skritih plasti	Število vhodnih nevronov	Število izhodnih nevronov	WER [%]
p-norm	2	1000	200	19,85
p-norm	4	1000	200	19,22
p-norm z ozkim grlom	2	1000	200	19,73
p-norm z ozkim grlom	4	1000	200	19,04
p-norm z ozkim grlom	4	2000	400	19,36
p-norm z ansamblom	4	1000	200	19,54
p-norm z ansamblom	4	2000	400	19,59

Osnovna aktivacijska funkcija p-norm doseže najboljši rezultat, ko uporabimo 1000 nevronov na vhodu in 200 na izhodu s štirimi plastmi. Sistem, ki ima samo dve skriti plasti, doseže za 0,63 % slabši WER. Rezultat nekoliko izboljšamo v kombinaciji z ozkim grlom, kjer uporabimo 1000 nevronov na vhodu, 200 na izhodu, implementirane pa so bile 4 skrite plasti. V kombinaciji z ozkim grlom dosežemo nato tretji najboljši WER 19,36 %, ki je zgolj za 0,14% slabši od arhitekture s samo p-norm aktivacijsko funkcijo in 0,32 % slabši od najboljšega rezultata. Najslabši rezultat dobimo v kombinaciji aktivacijske

funkcije p-norm in ozkega grla z dvema skritima plastema, 1000 vhodnimi in 200 izhodnimi nevroni. V primerjavi z najboljšim rezultatom, doseženim zgolj s p-norm aktivacijsko funkcijo razpoznave govora, je za 0,51 % slabša in za 0,69 % slabša v primerjavi z najboljšim rezultatom aktivacijske funkcije p-norm v kombinaciji z ozkim grlom. Pri aktivacijski funkciji p-norm z ansamblom dosežemo boljši rezultat, če izberemo manj nevronov, in sicer 1000 na vhodu in 200 na izhodu. Dobljeni WER je 19,54 % in je za samo 0,05 % slabši v primerjavi z enako kombinacijo z več nevroni na vhodu in izhodu. Od najboljšega rezultata s samo p-norm aktivacijsko funkcijo se razlikuje za 0,32 % in 0,50 % od najboljšega dosežena rezultata.

V drugem 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, ki so ga Žgank in sodelavci objavili leta 2014 in je dosegel najboljši WER 26,81 %. Rezultati napake razpoznavanja besed s trigramskim jezikovnim modelom in slovarjem besed z velikostjo 64.000 so predstavljeni v Preglednici 2.

Preglednica 2: *Rezultati razpoznavanja govora s trigramskim 64.000 jezikovnim modelom*

Sistem	**WER [%]**
netransformiran HMM	26,48
transformiran HMM	24,28
DNN s p-norm	19,22
DNN s p-norm in z ozkim grlom	19,04
DNN s p-norm ansamblom	19,54
DNN s tanh	18,76
DNN s tanh in z ozkim grlom	23,33

Izhodiščna primerjava akustičnih modelov HMM kaže, da je prehod na novo ogrodje za avtomatsko razpoznavanje govora potekal brez težav, saj smo dosegli zelo primerljiv WER (s 26,81 % na 26,48 %). Osnovne akustične modele HMM je možno dodatno nadgraditi z metodama od govorca neodvisne transformacije značilk z uporabo LDA (angl. Linear Discriminant Analysis) in MLLT (angl. Maximum Likelihood Linear Transform) (Gales, 1999), kar izboljša rezultat s 26,48 % na 24,28 %. Vendar je to izboljšanje relativno

omejeno v primerjavi z možnostmi, ki jih v ustreznih pogojih omogoča globoko učenje. Najboljši rezultat dobimo z aktivacijsko funkcijo tanh, kjer WER znaša 18,76 %. V kombinaciji z ozkim grlom se razpoznavanje govora poslabša za 4,57 %. Kombinacija z ozkim grlom je nekoliko doprinesla pri razpoznavanju z aktivacijo funkcijo p-norm, kjer je rezultat s samo aktivacijsko funkcijo izboljšala za 0,18 %. Aktivacijska funkcija p-norm v kombinaciji z ansamblom ne prinese izboljšanja, saj je rezultat za 0,32 % slabši v primerjavi s samo aktivacijsko funkcijo p-norm. Prehod na globoke nevronske mreže za akustično modeliranje izboljša napako razpoznavanja besed na 18,76 %, kar predstavlja statistično pomembno razliko. Pri tem je treba posebej izpostaviti, da je količina govornega učnega materiala relativno omejena z vidika metod globokega učenja. Za učenje akustičnega modela z aktivacijsko funkcijo tanh smo na grafični kartici z NVIDIA grafičnim procesorjem V100 potrebovali 15,5 ur. Čas dekodiranja testnega nabora pa je trajal 22 minut, tako da je bil faktor realnega časa xRT približno 0,12.

V drugem koraku smo izvedli vrednotenje, kako vpliva na rezultate izboljšani jezikovni model z bistveno večjim slovarjem besed. Prehod s 64.000 besed na 250.000 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 64.000. Se pa poveča perpleksnost takšnega jezikovnega modela. Rezultati razpoznavanja govora z akustičnimi modeli DNN in obema trigramskima jezikovnima modeloma so predstavljeni v Preglednici 3.

Preglednica 3: *Rezultati razpoznavanja govora z akustičnimi modeli DNN z različnima trigramskima jezikovnima modeloma*

Jezikovni model	WER [%]
64.000-3g	19,22
250.000-3g	15,17

Tudi v scenariju razpoznavanja govora s slovarjem besed z velikostjo 250.000 je prišlo do znatnega zmanjšanja napake razpoznavanja besed, saj je WER znašal 15,17 %. S povečanjem slovarja razpoznavalnika govora smo tako izboljšali delovanje za 4,05 %, 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 idr., 2019. Dosegli so WER 27,16 % na bazi GOS VideoLectures 2.0. Pri gradnji akustičnega modela so dodali tudi učenje s prilagajanjem govorcu (angl. speaker adaptive training), ki smo ga mi v gradnji izpustili. Modele GMM-HMM so nato uporabili kot osnovo za učenje modela DNN-HMM. Uporabili so arhitekturi TDNN in LSTM, preizkušali pa so več različnih konfiguracij mrež, kjer so različno povezovali plasti in spreminjali število skritih plasti. Zaradi uporabe različnih govornih in jezikovnih virov doseženi rezultati sicer niso neposredno primerljivi med seboj.

Najboljši doseženi rezultat razpoznavanja govora z jezikovnim modelom 250.000 3g je že primerljiv oziroma se je zelo približal rezultatom razpoznavanja govora v domeni televizijskih oddaj v nekaterih drugih jezikih. Avtorji v (Lleida idr., 2019) poročajo, da je na tekmovanju Albayzin RTVE 2018 Challenge za španščino najboljši sistem dosegel WER 16,45 %. Pri tem so uporabljali učni nabor posnetkov v dolžini več kot 200 ur.

V naslednjem koraku smo primerjali rezultate, ki smo jih pridobili s testnimi množicami, kjer smo uporabili dodatno izgubno kodiranje zvočnih zapisov.

Preglednica 4: *Rezultati razpoznavanja govora z vplivom izgubnih kodekov*

Kodek	WER [%]
MP3-64 kbit/s	19,21
MP3-128 kbit/s	19,10
AAC-64 kbit/s	18,96
AAC-128 kbit	18,84
MP3+AAC- 128kbit/s	19,47

Najboljši rezultat dobimo z izgubnim kodekom AAC, ki prinaša 0,08 % slabši rezultat glede na rezultat, ki smo ga dobili s posnetki v formatu WAV. Slabše se odreže kodek MP3, ki ima za 0,45 % slabši rezultat pri bitni hitrosti 64 kbit/s in za 0,34 % slabši rezultat pri 128 kbit/s. Manjša bitna hitrost poslabša rezultat za približno 0,1 %. Najslabše rezultate prinese transkodiranje, rezultat se poslabša za 0,71 %. Kodiranje z izgubnimi kodeki ne prinaša velikega poslabšanja rezultatov razpoznavanja govora. Na njihovi podlagi lahko predpostavimo, da bi takšen razpoznavalnik govora učinkovito deloval tudi

s posnetki, ki uporabljajo izgubne kodeke. Podobno kakor je bilo prikazano v članku (Pollak in Behunek, 2011), kjer so primerjali razpoznavanje govora z izgubnim kodekom MP3 pri različnih bitnih hitrostih, lahko opazimo, da je razpoznavalnik govora sposoben učinkoviteje razpoznavati posnetke, kadar je na voljo govor, kodiran z višjo bitno hitrostjo.

V nadaljevanju poglavja bomo podrobneje predstavili analizo doseženih rezultatov razpoznavanja govora. Tukaj smo uporabili akustične modele brez dodatne nadgradnje v obliki transformacije značilk. Odgovoriti poskušamo na vprašanje, kako različni faktorji vplivajo na WER. V to skupino sodijo delež besed izven slovarja, pregibna oblika besed, akustično ozadje in način govora.

Referenčne transkripcije in rezultate razpoznavanja smo oblikoslovno označili ter lematizirali z označevalnikom slovenskega jezika Obeliks (Grčar idr., 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. Za lematizacijo smo se odločili, ker pravilno razpoznana lema poda več informacij kot pa pravilno razpoznani koren besede ali uporaba deleža napačno razpoznanih znakov. V primeru pravilno razpoznane leme lahko predvidevamo, da se v večji meri ohrani pomen kot pa v primeru pravilno razpoznanega korena besede. S tem želimo pridobiti boljšo oceno, ali bi bralec avtomatske transkripcije lahko pravilno razumel pomen stavka in opazil le slovnično napako, medtem ko bi se pri pravilno razpoznanem korenu besede spremenil pomen stavka. Ta razlika je še bolj očitna v primeru uporabe deleža napačno razpoznanih znakov, saj lahko en napačni znak spremeni pomen stavka. S pomočjo oblikoslovnih oznak pa smo nato še napake v besedni obliki razdelili po besednih vrstah.

V Preglednici 5 so predstavljeni podrobnejši rezultati. Pri izhodiščnih rezultatih za HMM in DNN s 64.000 besedami v slovarju je prišlo do manjšega odstopanje v WER v primerjavi s prejšnjimi rezultati. Razlog za to odstopanje je uporaba drugega orodja za analizo rezultatov, ki nekoliko drugače poravna

rezultate razpoznavanja govora z referenčnimi transkripcijami. 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 4,29 %. To razliko bo v prihodnosti treba še podrobneje analizirati. Večja izboljšanja vidimo v razredih F1, F3 in FX pri prehodu s sistema HMM na DNN ter pri razredu F2 pri prehodu na večji slovar, ki pa predstavlja le zelo majhen del testne množice. Medtem ko za bran studijski govor dosegamo WER 7,83 %, sprememba na spontani govor ali dodajanje akustičnega ozadja poslabša rezultate v rangu 10 do 21 %. Pri tem je pričakovano poslabšanje večje, če je v ozadju dodana glasba.

Preglednica 5: *Podrobnejša predstavitev rezultatov razpoznavanja po F-razredih in spolu ter rezultati pravilnosti razpoznave lem*

Sistem	HMM 64.000-3g	DNN 64.000-3g	DNN 250.000-3g
WER [%]	24,33	18,82	15,17
WER – F0 [%]	14,63	11,46	7,83
WER – F1 [%]	31,57	24,27	20,99
WER – F2 [%]	58,47	39,83	38,14
WER – F3 [%]	33,43	25,83	21,01
WER – F4 [%]	27,14	21,10	17,28
WER – FX [%]	31,95	24,15	21,45
WER – Moški [%]	26,16	20,74	17,06
WER – Ženske [%]	21,95	16,43	12,77
LER [%]	23,33	17,70	14,06
WER – LER	1,00	1,12	1,11

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 bila v nekaterih primerih 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 se 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 okoli 1 %.

V nadaljevanju smo pregledali napake v besedni obliki pri isti lemi glede na besedno vrsto. Rezultati so podani v Preglednici 6. Podali smo le pregibne besedne vrste (brez zaimkov). Primerjamo sistema HMM 64.000-3g in DNN 250000-3g. 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 je približno dvakrat manjše od relativnega izboljšanja skupnega rezultata.

Preglednica 6: *Napačno razpoznane besedne oblike glede na besedno vrsto*

Besedna vrsta	Št. napak v HMM 64.000-3g	Št. napak v DNN 250.000-3g	Relativna izboljšava [%]
Samostalnik	309	265	14,2
Pridevnik	155	112	27,7
Glagol	148	135	8,8
Številnik	16	10	37,5
Prislov	0	1	-
SKUPAJ	628	522	16,9

Rezultati kažejo na to, da sistem s povečanim slovarjem in uporabo globokih nevronskih mrež pomembno zmanjša skupni delež napak razpoznavanja. Vidimo lahko, da je relativno zmanjšanje napak zaradi pregibnosti besed manjše glede na skupno zmanjšanje. V sistemu DNN 250.000-3g 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 250.000-3g.

Doseženi rezultati kažejo, da je z obstoječimi slovenskimi govornimi viri možno učinkovito graditi razpoznavalnike govora za domeno dnevnoinformativnih oddaj, če govorimo o preprostejših akustičnih pogojih. Kakor hitro pa dodamo zahtevnejše akustične pogoje, se rezultati poslabšajo. S tega vidika je pomembno delo na povečevanju razpoložljivih govornih virov za slovenski jezik. Z vidika visoke pregibnosti slovenskega jezika se je pokazalo, da lahko to lastnost učinkovito naslovimo z zniževanjem deleža besed izven slovarja. Na takšen način lahko modeliramo večino besed, težavna kategorija pa ostajajo kratke besede, ki so si akustično podobne. Za izboljšano akustično modeliranje v takšnih primerih pa je ponovno neobhodno potrebno več učnega govornega materiala. Pristop z zmanjševanjem deleža besed izven slovarja kaže, da je za doseganje primerljivih rezultatov razpoznavanja govora z jeziki, kot je angleščina, potreben za 3- do 5-krat večji slovar razpoznavalnika govora.

7 SKLEP

V članku smo predstavili sistem za avtomatsko razpoznavanje slovenskega govora v domeni televizijskih oddaj. Najboljši doseženi rezultat deleža napake razpoznavanja besed je znašal 15,17 %. Takšen sistem je po svojih rezultatih razpoznavanja govora že primerljiv z nekaterimi rezultati, doseženimi za druge jezike. 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. V prihodnjem delu se je tako smiselno osredotočiti na povečanje gradiva za učenje akustičnih modelov in s tem povezane spremembe v arhitekturi takšnih modelov.

Zahvala

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

Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069. Raziskovalno delo je bilo delno opravljeno v okviru projekta RSDO – Razvoj slovenščine v digitalnem okolju. Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

LITERATURA

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

Dobrišek, S., Gros, J., Mihelič, F., & Pavešić, N. (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.

Dobrišek, S., & Mihelič, F. (2010). Zmanjševanje odvečnosti končnih pretvornikov za učinkovito gradnjo razpoznavalnikov slovenskega govora z velikim besednjakom. V *Jezikovne tehnologije: zbornik 13. mednarodne multikonference, Informacijska družba IS* (str. 24–27).

Dobrišek, S., Žganec Gros, J., Žibert, J., Mihelič, F., & Pavešić, N. (2017). Speech Database of Spoken Flight Information Enquiries SOFES 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1125

ELRA. (2015). Pridobljeno s http://www.elra.info

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

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*,

Ljubljana, Slovenija (str. 89–94). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt12/JezikovneTehnologije2012.pdf

Imperl, B., Kačič, Z., & Horvat, B. (1996). Razpoznavanje osamljenih besed s polzveznimi Prikritimi modeli Markova. V *Zbornik pete Elektrotehniške in računalniške konference ERK* (str. B/231–234).

Imperl, B., & Kačič, Z (1999). Connected digits and natural numbers recognition for the telephone multilingual speech dialog systems. *V Proceedings of the 4th international workshop on Electronics, control, measurement and signals ECMS* (str. 164–167).

Ipšić, I., Mihelič, F., Dobrišek, S., Žganec Gros, J., & Pavešić, N. (1999). A Slovenian spoken dialog system for air flight inquiries. V *Eurospeech '99: proceedings, 6th European Conference on Speech Communication and Technology* (str. 2659–2662).

Kačič, Z., Horvat, B., & Greif, Š. (1988). Man-machine communication: speaker-independent speech recognition. *Informatica: an international journal of computing and informatics, (12)*1, 6–12.

Kaiser, J., & Kačič, Z. (1997). SpeechDat (II) Slovenian Database for the Fixed Telephone Network. Maribor, Slovenia: University of Maribor.

Kaiser, J., Sepesy Maučec, M., Kačič, Z., & Horvat, B. (2000). Razpoznavanje tekočega slovenskega govora z velikim slovarjem. V T. Erjavec in J. Gros (ur.), *Jezikovne tehnologije* (str. 39–44). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt00/zbornik/sdjt00-Kaiser06.pdf

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M., & De Prada, A. (2019). Albayzin 2018 evaluation: the iberspeech-RTVE challenge on speech technologies for spanish broadcast media. *Applied Sciences*, *9*(24), 5412.

Mihelič, F., Ipšić, I., Dobrišek, S., & Pavešić, N. (1992). Feature representations and classification procedures for Slovene phoneme recognition. *Pattern recognition letters*, *13*(12), 879–891.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A 463 systematic review. *IEEE Access* 2019, 7, 19143–19165.

Nouza, J., Safarik, R., & Cerva, P. (2016). ASR for South Slavic Languages Developed in Almost Automated Way. V *Interspeech* (str. 3868–3872).

Pollak, P., & Behunek, M. (2011). Accuracy of MP3 speech recognition under real-word conditions: Experimental study. V *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (str. 1–6). IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N.,..., Silovsky, J. (2011). The Kaldi speech recognition toolkit. V *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

RSDO. (b. d.). Pridobljeno s https://www.cjvt.si/rsdo/

Schwartz, R., Jin, H., Kubala, F., & Matsoukas, S. (1997). Modeling Those F-Conditions – or not. V *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. SRILM – an extensible language modeling toolkit. V *International Conference on Speech and Language Processing* (str. 901–904).

Ulčar, M., Dobrišek, S., & Robnik-Šikonja, M. (2019). Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika*. *27,* 3.

Verdonik, D., Kosem, I., Vitez, A., Krek, S., & Stabej, M. (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.

Verdonik, D., Potočnik, T., Sepesy Maučec, M., & Erjavec T. (2017). Spoken corpus Gos VideoLectures 2.0 (transcription). Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Pridobljeno s http://hdl.handle.net/11356/1222

Verdonik, D. (2018). Korpus in baza Gos Videolectures. V D. Fišer in A. Pančur (ur.), *Zbornik 11. konference Jezikovne tehnologije in digitalna humanistika* (str. 265–268). Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. Pridobljeno s http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf

Zhang X., Trmal, J., Povey, D., & Khudanpur, S. (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.

Zorrilla, A. L., Dugan, N., Torres, M. I., Glackin, C., Chollet, G., & Cannings, N. (2016). Some asr experiments using deep neural networks on spanish databases. *Advances in Speech and Language Technologies for Iberian Languages*. IberSPEECH.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2013). Spoken corpus Gos 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1040

Žgank, A., Kačič, Z., & Horvat, B. (2002). Preliminary evaluation of Slovenian mobile database PoliDat. V *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'02).

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., ..., Horvat, B. (2004). Acquisition and annotation of Slovenian broadcast news database. V *Fourth international conference on language resources and evaluation*, *LREC 2004* (str. 2103–2106). Lizbona, Portugalska. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2004/pdf/123.pdf

Žgank, A., Rotovnik, T., Grašič, M., Kos, M., Vlaj, D., & Kačič, Z. (2006). Sloparl-Slovenian parliamentary speech and text corpus for large vocabulary continuous speech recognition. V *Ninth International Conference on Spoken Language Processing*. Pridobljeno s http://dblp.uni-trier.de/db/conf/interspeech/interspeech2006.html#ZgankRGKVK06

Žgank, A., Rotovnik, T., Sepesy Maučec, M., & Kačič, Z. (2006). Osnovna zgradba razpoznavalnika slovenskega tekočega govora UMB Broadcast News. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS* (str. 99–118). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/is-ltc06/proc/

Žgank, A., & Sepesy Maučec, M. (2010). Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije 2010* (28–31). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt10/JezikovneTehnologije2010.pdf

Žgank, A., Donaj, G., & Sepesy Maučec, M. (2014). Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov. V V T. Erjavec in J. Žganec Gros (ur.) *Zbornik 9. konference Jezikovne tehnologije, Informacijska družba IS* (str. 147–150). Ljubljana: Institut

Jožef Stefan. Pridobljeno s http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_(LT).pdf

Žgank, A., Sepesy Maučec, M., & Verdonik, D. (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).

Žibert, J., Mihelič, F., & Dobrišek, S. (2000). Avtomatično podnaslavljanje vremenskih napovedi. V B. Zajc (ur.), *Zbornik devete Elektrotehniške in računalniške konference, Portorož, Slovenija, 21.–23. september 2000* (str. 165–168).

Žibert, J., Martinčić-Ipšić, S., Ipšić, I., & Mihelič, F. (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.

Žibert, J., & Mihelič, F. (2004). Development of Slovenian broadcast news speech database. V *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (str. 2095–2098). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2004/pdf/98.pdf

SLOVENIAN AUTOMATIC SPEECH RECOGNITION FOR BROADCAST NEWS

In speech and language technologies, automatic speech recognition is one of the key building blocks. In this article, we will explain the development of an automatic recognizer of Slovenian speech for the domain of daily news broadcasts. The architecture of the system is based on a deep neural net. Considering the available speech sources, we performed modeling with various activation functions. In the development of speech recognition, we also checked the impact of lossy speech codecs on speech recognition results. We used the UBM BNSI Broadcast News and IETK-TV databases to train the speech recognizer. The total amount of voice recordings was 66 hours. In parallel with the deep neural networks, we increased the speech recognition dictionary, which amounted to 250,000 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 15.17%. While evaluating the results, we also performed a more detailed analysis of speech recognition errors based on lemmas and F-conditions, which to some extent show the complexity of the Slovenian language for such scenarios of technology use.

Keywords: automatic speech recognition, characteristics of Slovenian language, broadcast news, deep neural networks, lossy speech codecs