Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora

Lucija Gril,* Simon Dobrišek, ‡ Andrej Žgank*

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
lucija.gril@um.si, andrej.zgank@um.si
‡ Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
simon.dobrisek@fe.uni-lj.si

Povzetek

V članku je predstavljen sistem avtomatskega razpoznavanja govora za slovenski jezik. Za graditev akustičnih modelov smo uporabili dva različna jezikovna vira in dve različni osnovni akustični enoti pri zapisu slovarjev. Testiranje je potekalo na testni množici, ki je nastala znotraj projekta Razvoj slovenščine v digitalnem okolju in vsebuje malo manj kot 5 ur zvočnih posnetkov. Za graditev jezikovnih modelov smo uporabili hibridni pristop HMM-DNN. Za nevronske mreže smo uporabili dva tipa mrež, in sicer TDNN in LSTM. Najboljši rezultat WER je znašal 24,95 % in smo ga dosegli z arhitekturo TDNN in grafemskim slovarjem.

Acoustic modeling with various basic units for Slovenian automatic speech recognition

The article presents the automatic speech recognition system for the Slovenian language. We used two different language sources and lexicons based on two basic acoustic units. The system was tested by the test set containing a little less than 5 hours of sound recordings that developed by the RSDO project. We used the hybrid HMM-DNN approach to build language models. Two types of networks were used for neural networks, namely TDNN and LSTM. The best WER score was 24.95% and we achieved it with TDNN architecture and grapheme lexicon.

1. Uvod

Dandanes nas pametna okolja spremljajo že na vsakem koraku. Pametni telefoni, tablice, televizijski sprejemniki, ročne ure, naprave v gospodinjstvu itd. Vse te naprave so nagnjene k temu, da nam nudijo boljšo in preprostejšo uporabniško izkušnjo. Storitev, ki jih nudijo, je veliko in za vse je potrebno skrbno načrtovanje tako strojne kot tudi programske opreme. Ena izmed takšnih storitev je tudi avtomatsko razpoznavanje govora (angl. Automatic speech recognition – ASR). Če želimo razpoznavati govor, se je treba zavedati, da lahko programska oprema deluje brezhibno, vendar na uspešnost njenega delovanja vpliva še veliko drugih dejavnikov. Eden izmed njih je lahko na primer slab mikrofon, ki zajame veliko šuma in popači zvok ter tako degradira razpoznavanje govora. To posledično vodi tudi do slabše uporabniške izkušnje. Prav tako lahko do poslabšanja rezultatov pride, če je razpoznavalnik tekočega govora slabše zasnovan in nima optimalnih karakteristik. Zato je pomembno, da z eksperimenti preverjamo različne arhitekture in zasnove modelov avtomatskega razpoznavalnika govora.

Za razvoj razpoznavalnika govora potrebujemo veliko količino učnega gradiva za posamezen jezik. Za jezike z veliko govorci je takšnega dosegljivega gradiva praviloma veliko. Za jezike z manjšim številom govorcev, kamor lahko uvrščamo tudi slovenščino, pa takšnih virov ni dovolj za uporabo naprednih metod umetne inteligence, kot je na primer enovito učenje (ang. end to end) s konvolucijskimi mrežami. V zadnjem obdobju se pogosto uporablja tudi učenje s prenosom znanja, vendar za obe našteti metodi velja, da omogočata slabši nadzor nad modeliranjem v primerjavi s hibridnim pristopom, ki smo ga uporabili v tem članku. Praviloma hibridni pristop tudi dosega nekoliko boljše rezultate, kot pa druga dva pristopa. Za avtomatski razpoznavalnik govora potrebujemo govorne posnetke, ki jih spremljajo datoteke s transkripcijo, v katerih je zapis izgovorjenih besed. Hkrati potrebujemo besedilni korpus in slovar, s katerima se lahko razpoznavalnik govora nauči značilnosti besed in tudi njihovega kontekstnega uvrščanja.

Izgovorjene besede lahko v avtomatskem razpoznavalniku govora predstavimo z dvema različnima akustičnima enotama – s fonemi ali z grafemi. Fonemi so glasovne enote, ki predstavljajo izgovorjavo glasov v besedi. Fonemski zapis slovenske izgovorjave se v večini primerov razlikuje od grafemskega. Grafemi in fonemi se med seboj razlikujejo tudi v številu osnovnih enot. Grafem zapišemo z eno osnovno enoto, ki ustreza črki v besedi. Po drugi strani se lahko ista črka slika v več različnih fonemov, odvisno od konteksta, naglasa in mesta v besedi. Prav tako se lahko črka preslika v zaporedje dveh fonemov. Z mislijo na to lahko pri razpoznavalniku tvorimo slovarje, ki vsebujejo izgovorjene besede na dva načina, in sicer s fonemi ali grafemi. Izbira vrste slovarja razpoznavalnika govora neposredno vpliva na to, kakšna bo osnovna akustična enota. Izbira akustične enote vpliva tudi na zahtevnost in način priprave slovarjev, kompleksnost akustičnih modelov in prek tega na potreben pomnilnik in procesorske zmogljivosti za učenje in delovanje avtomatskega razpoznavalnika govora. Tvorjenje slovarja s fonemi je odvisno od jezika, ki ga želimo uporabiti. Za slovenski jezik je ta naloga razmeroma zapletena in kompleksa. Slovar se lahko tvori ročno, kar praviloma počnejo fonetiki ali slovenisti, ali avtomatsko. Pri avtomatskih postopkih pa se lahko zgodi, da je zapis besede fonetično napačen, kar se v kasnejših korakih odraža na neoptimalnem učenju in razpoznavanju govora. Priprava slovarja z grafemi je lažja, saj je pretvorba trivialna. Kateri pristop je primernejši, je odvisno tudi od količine učnih podatkov, ki jih uporabljamo, saj je pri slovarjih s fonemi, ki so sestavljeni iz več osnovnih enot, večji poudarek na

številski porazdelitvi glede na kategorijo. V okviru projekta Razvoj slovenščine v digitalnem okolju (RSDO, b. d.) trenutno vzporedno poteka graditev govorne baze in pa razvoj prvih verzij avtomatskega razpoznavalnika govora. Zato imamo trenutno še vedno na voljo dokaj omejeno količino transkribiranega slovenskega govora, kar je bil povod za uporabo grafemske akustične enote. Že v preteklosti se je tako za slovenski jezik (Žgank in Kačič, 2006) kot tudi za druge jezike (Killer et at., 2003) pokazalo, da je lahko v takšnih primerih uporaba grafemskih akustičnih enot dobra rešitev. Tako smo za cilj članka postavili primerjavo med fonemskimi in grafemskimi akustičnimi osnovnimi enotami v povezavi s trenutno razpoložljivimi govornimi viri.

V nadaljevanju članka najprej pregledamo, kaj je na področju modeliranja akustičnih osnovnih enot avtomatskega razpoznavanja govora že bilo izvedenega za slovenski jezik. V tretjem poglavju predstavimo, katere govorne in jezikovne vire smo uporabili pri zasnovi eksperimentov. Tvorjenje slovarjev in samodejno grafemsko-fonemsko pretvorbo na osnovi pravil predstavimo v četrtem poglavju. V petem poglavju predstavimo modeliranje akustičnih in jezikovnih modelov avtomatskega razpoznavalnika govora. Rezultati so predstavljeni in komentirani v šestem poglavju, ki mu sledi še zaključek.

2. Pregled sorodnih člankov

Avtomatski razpoznavalniki govora so kot svojo privzeto akustično osnovno enoto uporabljali foneme in njihove izpeljanke v obliki kontekstnega podaljševanja. Izhodišče je bilo, da gre pri avtomatskem razpoznavanju govora za pretvorbo iz govorjene v besedilno obliko, kar se tako sklada z izbiro osnovne akustične enote. Leta 2000 so Schillo in sodelavci predstavili prvi grafemski avtomatski razpoznavalnik govora, ki je z izbiro drugačne osnovne akustične enote kršil navedeno predpostavko. Sistem je za nemški jezik sicer dosegel slabše rezultate razpoznavanja govora kot fonemski sistem, vendar so bili naučeni grafemski modeli manjši.

Grafemi kot osnovne akustične enote postanejo hitro zanimivi tudi za večjezično in križnojezično razpoznavanje govora (Killer et at., 2003). V takšnih primerih je namreč možno združevati jezike brez podrobnega poznavanja fonetike vključenih jezikov. Osnovo pač predpostavlja zapisana črka. Uporabnost takšnega pristopa pride še dodatno do izraza pri križnojezičnem razpoznavanju govora, kjer so v ciljnem jeziku na voljo omejeni govorni viri. Uspešnost metode je v določeni meri odvisna tudi akustično-fonetične podobnosti med vključenimi jeziki.

Prve raziskave o uporabi grafemov kot osnovne akustične enote za križnojezično razpoznavanje slovenskega govora so predstavili Žgank in sodelavci (2005). Sledila je še uporaba grafemov za običajno enojezično avtomatsko razpoznavanje govora (Žgank in Kačič, 2006). Grafemi kot osnovne akustične enote so tako postali del standardne izbire za razpoznavanje slovenskega govora, še posebej v domeni dnevnoinformativnih oddaj (Gril et al., 2021). V kombinaciji s slovenskimi razpoznavalniki govora, ki so zasnovani na HMM akustičnih modelih ali na hibridni zasnovi HMM/DNN in imajo za učenje na voljo nekaj 10 ur transkribiranih govornih posnetkov, praviloma dosežejo boljše rezultate razpoznavanja govora. Predpostavimo sicer lahko, da se bo ta razlika manjšala, ko bo za slovenski jezik na voljo več ur transkribiranega govora. Z večanjem količine posnetkov namreč pridobimo na posamezno osnovno enoto tudi več vzorcev, kar izboljša možnost modeliranja akustičnih značilnosti in izboljša robustnost na potencialne napake, ki se lahko zgodijo zaradi avtomatske grafemsko-fonemske pretvorbe.

3. Govorni in jezikovni viri

Govorni in jezikovni viri so pri razpoznavalnikih govora ključna komponenta. Za govorne posnetke smo uporabili korpuse Gos 1.0 (Zwitter Vitez et al., 2013), Sofes (Dobrišek et al., 2017) in delovno različico testnega seta nastajajoče govorne baze RSDO (trenutna delovna različica je 2.0, ki ne vsebuje več črkovanja). Korpusa Gos in Sofes smo uporabili za učno in razvojno množico, medtem ko smo testni korpus 2.0 projekta RSDO uporabili za vrednotenje rezultatov. Za slovarje smo uporabili prostodostopni vir Sloleks 2.0 (Dobrovoljc et al., 2019) in trenutno verzijo slovarja, ki je nastala v projektu RSDO. Za besedilni korpus smo uporabili prostodostopni besedilni vir ccGigafida 1.0 (Logar et al., 2013).

Korpus Gos vsebuje 120 ur posnetkov. Posnetki zajemajo različne zvrsti, npr. televizijske oddaje, predavanja, pouk, zasebne pogovore … Ves govor je transkribiran v dveh različicah, in sicer v pogovorni in standardizirani različici. Posnetki zajemajo 1526 različnih govorcev. Govorni korpus Sofes vsebuje 9 ur in 52 minut posnetkov, ki zajemajo govor 134 različnih govorcev. Posnetki vsebujejo poizvedovanja po letalskih informacijah v slovenskem jeziku. Pri korpusu Sofes najdemo transkripcije v fonetičnem zapisu in standardiziranem zapisu govora. V testnem setu 2.0 RSDO je za 4 ure in 47 minut posnetkov. Korpus se od različice 1.0 razlikuje po tem, da ne vsebuje posnetkov črkovanja, kar znaša okoli 19 minut govora. Črkovanje smo iz splošnega testnega nabora izločili, saj je za njegovo učinkovito razpoznavanje treba uporabiti drugačne pristope. Testna množica RSDO zajema bran, javni, nejavni govor in posnetke državnega zbora. V posnetkih se pojavi 63 različnih govorcev. Tudi pri korpusu RSDO imamo dva različna zapisa govora, ki sta nastala na osnovi enakih priporočil kot v korpusu Gos.

Vir Sloleks 2.0 je leksikon, ki vsebuje okoli 2.792.000 posameznih besednih oblik. Vsak vnos vsebuje podatke o besedi (v katero besedno vrsto sodi in kakšne so njene slovnične lastnosti). Zapisane so tudi vse pregibne oblike za posamezno besedo. Slovenščina je pregiben jezik in zato je takšnih oblik zelo veliko. V različici 2.0 je označeno tudi mesto naglasa in zapis v mednarodni fonetični pisavi (IPA).

V našem primeru smo Sloleks 2.0 uporabili za tvorjenje fonetičnega slovarja avtomatskega razpoznavalnika govora. V takšnem slovarju potrebujemo besede in njihovo izgovorjavo s fonemi. Sloleks 2.0 smo s pomočjo postopka, ki so ga uporabili Ulčar in drugi (2019), pretvorili v obliko, ki je ustrezna za avtomatski razpoznavalnik govora.

Besedilni korpus CcGigafida vsebuje nekaj čez 103.000.000 besed in je javno dostopni del korpusa Gigafida, ki ga je možno uporabljati pod licenco Creative Commons. Besedilo vsebuje informacije o virih časopisov, revij, leta izdaj, vrsti besedil, naslovih, o avtorjih besedil. Korpus je označen z morfoskladenjskimi opisi in lemami.

Besedilni korpus ccGigafida smo uporabili za jezikovno modeliranje avtomatskega razpoznavalnika

govora. Zaradi pravilne obdelave smo iz korpusa izbrisali prazne vrstice in večkratne presledke. Odstranili smo tudi ločila, da je bilo besedilo v skladu z običajno obliko v sistemu za razpoznavanje govora.

4. Tvorjenje slovarjev za razpoznavalnik govora

Tvorjenje fonetičnih slovarjev, ki so potrebni za graditev hibridnih arhitektur avtomatskih razpoznavalnikov govora, temelji tako na uporabi obstoječih razpoložljivih leksikonov, ki so navadno ročno preverjeni in že vsebujejo fonetične prepise besed, kot tudi na uporabi samodejnih grafemsko-fonemskih pretvornikov, ki se uporabljajo za t. i. izvenslovarske besede, ki jih predvideva jezikovni model razpoznavalnika govora, niso pa še vključene v obstoječe leksikone.

Tvorjenje slovarja za prvo različico avtomatskega razpoznavalnika govora (»Rezultat R2.2.7: Orodje za grafemsko fonemsko pretvorbo – verzija 2«, 2022) , ki je bil razvit v okviru projekta RSDO in bo predstavljen v naslednjih poglavjih, je primarno temeljilo na uporabi že omenjenega leksikona Sloleks 2.0 ter ročno urejenega in preverjenega slovarja izgovorjav, ki je vključen v govorni korpus Sofes. Za vse besede, ki se pojavljajo v normiranih besednih prepisih vseh zvočnih govornih posnetkov, ki so se uporabili za tvorjenje akustičnega modela razpoznavalnika govora, ter za vse besede, ki se pojavljajo v normiranem besedilnem korpusu, ki se je uporabil za tvorjenje njegovega jezikovnega modela, smo najprej pogledali v leksikon Sloleks 2.0 in ročno urejen slovar govornega korpusa Sofes, če ta morda vsebujeta obravnavano besedo. Če je bila beseda v tem leksikonu oziroma slovarju vsebovana, se je njen fonetični prepis samo prenesel v slovar razpoznavalnika govora. Če obravnavana beseda v leksikonu Sloleks 2.0 oziroma slovarju Sofes ni bila vsebovana, pa se je njen fonetični prepis pridobilo z uporabo prve različice samodejnega grafemsko-fonemskega pretvornika, ki je bil razvit v okviru projekta RSDO in je v grobem opisan v nadaljevanju. Pri tvorjenju slovarja za predstavljeni razpoznavalnik govora se je samodejno moralo pretvoriti kar več kot 58 odstotkov vseh besed, ki so bile predvidene za razpoznavalnik govora. Pravilnost samodejne pretvorbe pa pri prvi različici razpoznavalnika govora še ni bila natančno preverjena in ovrednotena.

4.1. Samodejna grafemsko-fonemska pretvorba na osnovi pravil

Prva različica samodejnega grafemsko-fonemskega pretvornika, ki je bil razvit v okviru projekta RSDO in se je uporabil za tvorjenje slovarja razpoznavalnika govora, je temeljila na uporabi množice kontekstno odvisnih fonetičnih pravil, ki so bila določena na osnovi statističnih analiz in obstoječega jezikoslovnega in glasoslovnega poznavanja fonetičnih značilnosti slovenskega govorjenega jezika. Upoštevana kontekstno odvisna pravila so temeljila predvsem na upoštevanju mesta naglasa v danih besedah.

Mesto naglasa v besedi na splošno določa zlog, na katerem ima beseda jakostno ali tonsko izraženo slušno zaznavno izrazitost (Toporišič, 1992). Pomembna značilnost slovenskega jezika je, da se mesto naglasa pojavlja na prvem, zadnjem, predzadnjem ali tudi predpredzadnjem zlogu. Poleg tega pa lahko imajo posamezne besede tudi več mest naglasa. Mesto naglasa je določeno za vsako besedo posebej in se ga je med različnimi generacijami govorcev slovenskega govorjenega jezika zgodovinsko prenašalo z učenjem jezika in govornim sporazumevanjem. Kljub različnim mestom naglasa, ki so se z razvojem jezika in v različnih narečnih jezikovnih skupinah tudi spreminjala, pa je vendarle možno opredeliti določena pravila, ki v pretežni meri določajo mesto naglasa v besedah (Toporišič, 1991). Ta pravila so bila v glavnem upoštevana in uporabljena za samodejno določanje mesta naglasa v danih besedah. Pravila temeljijo na upoštevanju seznamov predpon, pripon, začetnic in končnic, ki se pojavljajo v slovenskih besedah in značilno določajo mesto naglasa v dani besedi. Pravila so bil določena na podoben način, kot je bilo to izvedeno pri razvoju sistema za samodejno tvorjenje umetnega slovenskega govora (Gros, 1997).

Uporabljena pravila sicer ne pokrijejo vseh trenutno uporabljanih slovenskih besed. Zato se je na osnovi dodatne statistične analize mest naglasov pri najbolj pogostih slovenskih besedah določilo še dodatna pravila za določitev najbolj verjetnega mesta naglasa v danih besedah. Ta pristop je do določene mere možno tolmačiti tudi kot izvajanje strojnega učenja iz podatkov.

Grafemski zapisi vhodnih besed se v razvitem pretvorniku z uporabo pravil pretvarjajo po vrsti, od leve proti desni. Pravila se v pretvorniku preverjajo in upoštevajo po danem vrstnem redu, zato si morajo slediti tako, da so na začetku seznama pravil za posamezen grafem najprej tista, ki opisujejo posebne primere pretvorb, sledijo pa jim bolj splošna pravila.

Razviti grafemsko-fonemski pretvornik na svojem vhodu predvideva besede, ki so že podane v normirani polni besedni obliki. Števila, števniki, denarne enote, okrajšave in drugi posebni zapisi morajo tako biti podani v polni besedni obliki. Za to je bilo poskrbljeno z normalizacijo besednih prepisov govornih posnetkov, ki so se uporabljali za tvorjenje akustičnega modela razpoznavalnika govora, in tudi besedil iz besedilnega korpusa, ki so se uporabljala za tvorjenje jezikovnega modela razpoznavalnika govora.

Izhodni nabor fonemskih različic je glede na samodejno določanje in upoštevanje mesta naglasa omogočal tudi ločevanje med dolgimi in kratkimi samoglasniki. Pri tvorjenju slovarja za razpoznavalnik govora pa se to ločevanje ni upoštevalo, ker se pri tvorjenju akustičnih modelov razpoznavalnikov govora samoglasnikov navadno ne ločuje na kratke in dolge, ker dolžina samoglasnikov nima osnovne pomensko razločevalne vloge pri razpoznavanju besed (Ulčar, 2019).

5. Arhitektura avtomatskega razpoznavalnika govora

Glede na razpoložljivo količino akustičnega učnega materiala, je bilo smiselno uporabiti hibridno arhitekturo avtomatskega razpoznavalnika govora, ki je v takšnih primerih praviloma učinkovitejša, kot so pa pristopi E2E.

Pri hibridnih sistemih avtomatskega razpoznavalnika govora lahko arhitekturno sestavo grobo razdelimo na dva dela, in sicer na akustični model in jezikovni model. Akustični model naučimo na osnovi vzorcev iz zvočnih posnetkov in njihovih ustreznih prepisov, jezikovni model pa glede na besedilni korpus. V nadaljevanju članka bomo podrobneje predstavili oba modela, za graditev katerih smo uporabili prostodostopno orodje Kaldi (Povey et al., 2011).

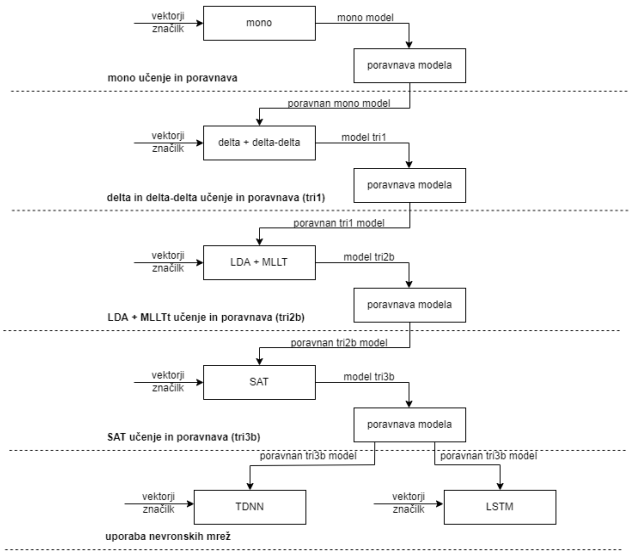
Za pripravo spremljajočih datotek, ki jih potrebujemo za graditev modela v orodju Kaldi, smo uporabili transkripcije govornih korpusov, ki so zapisane v obliki standardiziranega zapisa govora.

5.1. Akustično modeliranje

Za akustično modeliranje smo uporabili govorne baze Gos, Sofes in testno množico projekta RSDO. Zvočni posnetki govornih baz Gos in Sofes so bili v mono formatu in so bili zapisani v 16-bitnem zapisu. Frekvenca vzorčenja je bila 16 kHz. Posnetki testne množice projekta RSDO so imeli frekvenco vzorčenja 44,1 kHz, bitna hitrost in format pa sta bila enaka posnetkom v bazah Gos in Sofes. Orodje Kaldi za svoje delo potrebuje mono zvočne posnetke s frekvenco vzorčenja 16 kHz in 16-bitnim zapisom. Da lahko posnetke v orodju Kaldi procesiramo, moramo posnetke pretvoriti v ustrezni format. S prostodostopnim orodjem SoX smo posnetke pretvorili v mono zvočne posnetke, s frekvenco vzorčenja 16 kHz in 16-bitnim zapisom. Pretvarjanje posnetkov smo vključili v proces priprave potrebnih datotek za procesiranje v orodju Kaldi. S tem korakom smo se ognili ročnemu pretvarjanju vseh posnetkov.

Zvočne posnetke, ki so del govorne baze, smo spremenili v vektorje značilk. Na začetku posnetke razdelimo na okna dolžine 25 ms in jih nato transformiramo, da dobimo značilke MFCC (mel-frekvenčne kepstralne koeficiente). Za nadaljnje delo smo uporabili 12 MFCC koeficientov in energijo, nad katerimi smo izračunali še prvi in drugi časovni odvod. S prvim odvodom dobimo delta in z drugim delta-delta značilke. Nadaljevali smo z akustičnim modeliranjem, kjer smo v več fazah izvajali učenje novih akustičnih modelov in njihove poravnave.

Osnova akustičnega modeliranja avtomatskega razpoznavalnika govora so prikriti modeli Markova (angl. Hidden Markov Model – HMM). Z modeli HMM na osnovi vhodnih vektorjev značilk ocenjujemo verjetnosti hipotez izgovorjenega govora. Za to moramo poznati zapis fonemov v vsaki besedi. Takšne zapise imamo vnesene v fonetičnem slovarju, kjer je vsaka beseda predstavljena z nizom fonemov izgovorjene besede. Pri tem imamo lahko za posamezno besedo na voljo več izgovorjav, kar je odvisno od vključenega slovarja. Pri HMM modelih foneme predstavimo s skritimi stanji, model pa nato izračuna opazovana stanja s pomočjo Gaussovih porazdelitev, ki tvorijo hipoteze izgovorjene besede.



Slika 1: Postopek učenja akustičnega modela avtomatskega razpoznavalnika govora.

V naslednji fazi smo uporabili linearno diskriminančno analizo (angl. Linear discriminant analysis – LDA), s katero poiščemo linearno kombinacijo stanj. LDA vzame vektorje značilk in zgradi HMM stanja, vendar z manjšim prostorom značilke za vse podatke. LDA smo uporabili v kombinaciji z linearno transformacijo z največjo verjetnostjo (angl. Maximum Likelihood Linear Transform – MLLT), ki poenostavi računanje Gaussovih porazdelitev (Gales, 1999). MLLT vzame značilke iz LDA in izpelje edinstveno transformacijo za vsakega govorca. MLLT je prvi korak k normalizaciji govorcev, saj minimalizira razlike med govorci. Pri LDA in MLLT se uporabi prvih 13 značilk MFCC in vsako razdeli na 4 predhodna okna na levi in 4 naslednja okna na desni. To pomeni, da imamo končno dimenzijo značilk 117. Nato z LDA dimenzijo značilke omejimo na 40.

Za večjo natančnost avtomatskega razpoznavanja govora smo uporabili učenje s prilagajanjem govorcu (angl. Speaker Adaptive Training – SAT), ki za vsakega posameznega govorca izračuna parametre adaptacij glede na učne podatke tega govorca (Anastasakos et at., 1996).

Učenje akustičnega modela smo začeli z monofonskim akustičnim modelom in nadaljevali s trifonskim akustičnim modelom z delta in delta-delta (tri1) značilkami, trifonskim akustičnim modelom z LDA in MLLT (tri2b) ter na koncu še s trifonskimi akustičnimi modeli s SAT (tri3b). Postopek učenja je prikazan tudi s pomočjo diagrama, ki ga lahko vidimo na sliki 1.

V drugem delu graditve akustičnih modelov sledi prehod na globoke nevronske mreže. Nevronske mreže so sistemi, kjer algoritmi posnemajo delovanje nevronov v možganih. Sistem je sestavljen iz vhodnih, skritih in izhodnih plasti, ki so sestavljene iz enega ali več nevronov. Nevroni so med seboj povezani z relacijami, ki lahko potekajo naprej, nazaj ali naprej in nazaj. Na relacijah se uporabljajo uteži, s katerimi se izračunajo nova stanja.

Uporabili smo dva različna tipa nevronskih mrež, in sicer časovno zakasnjene nevronske mreže (angl. Time Delayed Neural Networks – TDNN) in nevronske mreže z

dolgim kratkoročnim spominom (angl. Long Short Term Memory – LSTM).

TDNN so nevronske mreže (Waibel, 1989), ki imajo več plasti. Začetne plasti se transformacije učijo bolj ozko, kasnejše pa imajo širši časovni kontekst. Za kontekstno modeliranje je treba zagotoviti, da vsaka nevronska celica poleg vhodne vrednosti, ki jo pridobi od aktivacijske funkcije oziroma iz nižje plasti, pridobi tudi informacijo o vzorcu izhodnih vrednosti in njihovega konteksta. Kar v primeru s časovnim signalom pomeni, da dobi vsaka nevronska celica na vhod informacijo o aktivacijskem vzorcu skozi čas od nižje ležečih plasti.

Nevronske mreže LSTM (Povey, 2018) vključujejo spominsko celico, ki ohrani informacijo dalj časa. Celica ima troje različnih vrat, in sicer vhodna, izhodna ter pozabljiva. Vhodna vrata uravnavajo količino podatkov prejšnjega vzorca, ki se bo shranila. Izhodna vrata določajo količino podatkov, ki se bo prenesla na naslednjo plast. Pozabljiva vrata pa regulirajo hitrost izgubljanja informacij v celici. Zaradi shranjevanja informacij so sistemi LSTM primerni za delo s časovnimi signali, saj se lahko pomembni dogodki zamaknejo. Modelu LSTM lahko rečemo tudi izboljšana ponavljajoča se nevronska mreža (angl. Recurrent Neural Network – RNN), saj je bila tako odpravljena težava izginjajočega gradienta (Hochreiter, 1991).

Arhitektura TDNN je sestavljena iz vhodnega nivoja, skritih nivojev in izhodnega nivoja. Vhodni nivo je dimenzije 40. Prvi skriti nivo mreže TDNN je bila mreža LDA z dimenzijo 40 in je bila polno povezana. Mreži LDA je sledilo še 8 polno povezanih mrež TDNN dimenzij 512. Na 8 nivojih mrež TDNN je bilo uporabljeno izpuščanje nevronov (angl. dropout). Mrežam TDNN sledita še dve vzporedni veji nivojev, in sicer verižna veja in veja xent. Verižna in xent veji sta sestavljeni iz dveh nivojev. Prva vzporedna nivoja tvorita mreži ReLU dimenzije 512. Mreži sta polno povezani in enako kakor mreže TDNN uporabljata izpuščanje nevronov. Mrežama ReLU sledita izhodna nivoja. Veji se razlikujeta po funkciji izgube. Verižna veja uporablja funkcijo logaritem verjetnosti pravilne sekvence fonemov oziroma grafemov, medtem ko veja xent za funkcijo izgube uporablja križno entropijo. Mreža TDNN je tako sestavljena iz 10 nivojev, pri katerih pa smo uporabili tudi časovno združevanje, kjer se na teh nivojih združijo informacije iz želenih časovnih oken glede na vhod. Časovno združevanje smo uporabili na nivoju LDA in 2., 4., 6., 7. ter 8. nivoju TDNN.

Učenje modelov TDNN je potekalo 7 epoh. Začetno efektivno stopnjo učenja (angl. initial effective lrate) smo nastavili na 0,0001 in končno (angl. final effective lrate) na 0,00001. Ostale vrednosti parametrov smo ohranili na privzetih vrednostih.

Tako kot arhitektura TDNN tudi LSTM vsebuje tri vrste nivojev. Prvi je vhodni in je enak vhodnemu nivoju arhitekture TDNN. Prav tako je tudi prvi skriti nivo arhitekture LSTM enak nivoju LSA, ki sestavlja arhitekturo TDNN. Naslednji štirje skriti nivoji so mreže LSTMP (angl. Long Short-Term Memory Projection) velikosti 1024. LSTMP je mreža LSTM, ki dodatno vsebuje še projekcijski nivo. V naši konfiguraciji arhitekture smo dimenzijo projekcijskega nivoja nastavili na 256. Skritim nivojem sledita dve veji izhodnih nivojev. Tudi tukaj se veji razlikujeta glede na funkcijo izgube tako kot pri arhitekturi TDNN.

Akustične modele LSTM za razpoznavanje govora smo učili s 4 epohami. Ostale vrednosti smo ohranili na privzetih, vključno z začetno in končno efektivno stopnjo učenja, ki sta bili nastavljeni na 0,001 in 0,0001.

V naslednjem poglavju bomo predstavili rezultate sistemov LSTM in TDNN za razpoznavanje govora. Ker je sistem TDNN dosegel boljše rezultate, smo del eksperimentov opazovali samo na sistemu TDNN.

5.2. Jezikovno modeliranje

Kot povezovalni člen med akustičnim in jezikovnim prostorom smo uporabili dva različna tipa slovarjev avtomatskega razpoznavalnika govora. Prvi tip uporabljenih slovarjev je bil fonemski slovar, kjer so besede zapisane s fonemi, in drugi tip, kjer smo namesto zapisa izgovorjene besede s fonemi uporabili zapis z grafemi. V tabeli 1 smo predstavili lastnosti slovarjev. Ena izmed lastnosti je tudi delež besede izven slovarja (angl. out of vocabulary – OOV), ki ga izračunamo kot:

$$OOV = \frac{\text{št. besed izven slovarja v testni množici}}{\text{št. vseh besed v slovarju}} \cdot 100 \quad (2)$$

Slovarji, ki smo jih uporabili, so večji, kakor tisti, ki so se uporabljali v prejšnjih razpoznavalnikih informativnih oddaj (Gril et at., 2021). Vrednosti OOV so zelo nizke in jih lahko enostavno zanemarimo.

Slovar	Tip slovarja	Št. besed	OOV [%]
Sloleks 2.0	fonemski	1.129.144	0,054
Sloleks 2.0	grafemski	931.848	0,065
RSDO	fonemski	1.440.070	0,008
RSDO	grafemski	1.440.070	0,008

Tabela 1: Lastnosti uporabljenih slovarjev.

Jezikovni model avtomatskega razpoznavalnika govora naučimo z besedilnim korpusom. Takšen model je sposoben predvidevati besedo, ki sledi, glede na predhodne besede v nizu. Jezikovni model ima tudi zmožnost kontekstnega uvrščanja, saj bo med besedami, ki imajo podobno izgovorjavo, izbral tisto, ki bo bolj smiselna glede na kontekst predhodno opazovanega zaporedja besed.

Jezikovni model smo naučili z uporabo orodja n-gram count, ki je del paketa SRILM (Stolcke, 2002). N-grami so v našem primeru nizi *n* besed v stavku. N-gram count glede na besedilni korpus generira n-grame in z njimi ocenjuje napovedne verjetnosti jezikovnega modela. Pri n-gram countu je treba določiti, do kakšne velikosti n-gramov želimo zgraditi model. Tako smo zgradili jezikovni model, ki je vseboval 1-grame, 2-grame in 3-grame.

6. Rezultati avtomatskega razpoznavanja govora

Uspešnost različnih verzij avtomatskega razpoznavalnika govora smo ovrednotili na testni množici 2.0 projekta RSDO. Za vrednotenje smo uporabili delež napačno razpoznanih besed (angl. Word Error Rate – WER). WER smo izračunali kot razmerje med številom vrinjenih, izbrisanih ter zamenjanih besed in med številom besed, ki so v referenčnem besedilu. To lahko zapišemo kot:

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

Kjer je *I* število vrinjenih besed (angl. insertions), *D* število izbrisanih besed (angl. deletions) in *S* število zamenjanih besed (angl. substitutions). Z *N* označimo število vseh besed v referenčnem besedilu testne množice. Razmerje nato pomnožimo s 100, saj WER praviloma podajamo v odstotkih.

Arhitektura	Slovar	Tip slovarja	WER [%]
LSTM	Sloleks 2.0	fonemski	38,70
TDNN	Sloleks 2.0	fonemski	27,19
TDNN	RSDO	fonemski	25,31
TDNN	Sloleks 2.0	grafemski	26,97
TDNN	RSDO	grafemski	24,95

Tabela 2: Rezultati razpoznavanja govora z različnimi vrstami vključenih metod in modelov.

Najprej poglejmo rezultate, ki smo jih dobili, ko smo vrednotili različna tipa arhitektur akustičnih modelov. Predstavljeni so v tabeli 2. Sistem LSTM se je izkazal za slabšega, saj je bil rezultat WER kar za 11,51 % slabši kot v primeru, ko smo uporabili sistem TDNN. Na osnovi tega rezultata smo kot nadaljnjo arhitekturo akustičnih modelov izbrali TDNN. Izhodiščni WER je bil 27,19 %. Ulčar in drugi (2019) so na podobnem sistemu dosegli malo slabši rezultat, vendar rezultati niso neposredno primerljivi, saj se je vrednotenje preverjalo na drugi testni množici. Primerjava s predhodnim podobnim ASR (Gril et at., 2021) kaže razliko v rezultatih. Avtorji so takrat dosegli 15,17 % WER, vendar z uporabo drugačnih govornih virov. Domena virov je bila v prejšnjem primeru omejena izključno na televizijske oddaje, res pa je, da so te lahko v nekaterih primerih, kot je na primer glasbeno ozadje govora, tudi dokaj kompleksne za avtomatsko razpoznavanje govora.

Za nadaljevanje razvoja sistema za razpoznavanje govora smo uporabili dva različna slovarja, in sicer slovar, ki je bil narejen na osnovi Sloleksa, in slovar, ki je bil pripravljen v sklopu projekta RSDO. V tabeli 2 lahko vidimo, da se rezultat vrednotenja z uporabo slovarja, ki je bil pripravljen pri projektu RSDO, izboljša za 1,88 %.

V zadnjem koraku smo primerjali med seboj še avtomatske razpoznavalnike govora, pri katerih smo z uporabo različnih tipov slovarja fonemsko osnovno akustično enoto zamenjali z grafemsko. Za avtomatski razpoznavalnik govora, pri katerem smo uporabili za osnovo Sloleks, je zamenjava fonemov z grafemi izboljšala rezultat za 0,22 %. Pri uporabi slovarja, ki je bil izdelan v okviru projektu RSDO, pa je zamenjava fonemov z grafemi WER izboljšala za 0,36 %. Rezultat s tem modelom in enotami je hkrati najboljši rezultat razpoznavanja govora, ki smo ga dosegli s predstavljenimi eksperimenti. Rezultat z grafemi je verjetno boljši zaradi omejene količine učnih podatkov in s tem tudi števila vzorcev na posamezno akustično enoto. Sklepamo lahko, da je teh bilo premalo za razpoznavo specifičnih akustičnih enot, ki so redkejše. Tako je razpoznavanje z grafemi, ki imajo manj akustičnih osnovnih enot, ker ne razlikujejo podvariant, delovalo bolje. Čeprav izboljšanje z grafemskim slovarjem ni posebej veliko, lahko pri tem tipu slovarja opozorimo na to, da je postopek priprave veliko preprostejši. Prednosti ima tudi pri uporabi, saj po velikosti zasede nekoliko manj pomnilniškega prostora, kar je posebej pomembno pri avtomatskih razpoznavalnikih govora z velikimi slovarji (angl. Large-Vocabulary Continuous Speech Recognition – LVCSR), kjer pri velikih datotekah hitro nastane ozko grlo. Dodatna prednost grafemskih akustičnih enot je tudi v tem, da lahko v praktični uporabi slovar avtomatskega razpoznavalnika govora nadgrajuje tudi laik.

7. Zaključek

V članku smo predstavili sistem za razpoznavo slovenskega govora. Za akustični model smo uporabili hibridni pristop HMM-DNN. Za napovedovanje skritih stanj v HMM smo uporabili dva tipa nevronskih mrež. Časovno zakasnjene nevronske mreže so se izkazale za boljši pristop kakor nevronske mreže z dolgim kratkoročnim spominom. Za tvorjenje slovarja smo uporabili dve osnovni akustični enoti. Grafemski modeli so v našem primeru dali boljše rezultate kakor fonemski. Uporabili smo novo testno množico, ki je nastala pri projektu RSDO. Najboljši delež napačno razpoznanih besed je bil 24,95 %. Rezultat je primerljiv tudi z rezultati drugih sistemov razpoznavanja govora. K dobremu rezultatu razpoznave prispeva velik slovar, ki je večji kakor pri primerljivih sistemih, in uporaba grafemov kot osnovne akustične enote. Sistemi z grafemi omogočajo enostavnejše tvorjenje slovarjev, enostavnejše je tudi nadgrajevanje takšnih slovarjev. Uporaba grafemov ima pozitivni učinek tudi pri uporabi sistemov, saj takšni modeli zavzemajo nekoliko manj pomnilniškega prostora.

Zahvala

Zahvaljujemo se avtorjem korpusa Gos 1.0, ki so nam omogočili njegovo uporabo za razvoj avtomatskega razpoznavalnika govora.

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.

8. Literatura

Tasos Anastasakos, John McDonough, Richard Schwartz in John Makhoul. 1996. A compact model for speaker-adaptive training. V: Proceedings ICSLP, str. 113–1140.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič in Nikola Pavešić. 2017. Speech Database of Spoken Flight Information Enquiries SOFES 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1125

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. Morphological lexicon Sloleks 2.0. Slovenian language resource repository CLARIN.SI, ISSN 2820–4042, http://hdl.handle.net/11356/1230

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

Jerneja Gros. 1997. *Samodejno tvorjenje govora iz besedil*. Doktorska disertacija. Fakulteta za elektrotehniko, Univerza v Ljubljani.

Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen Netzen. Dostopno na:

https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf (16. 5. 2022)

Mirjam Killer, Sebastian Stüker and Tanja Schultz. 2003. Grapheme based speech recognition. *Interspeech.*

Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar in Peter Holozan. 2013. Written corpus ccGigafida 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1035

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motliček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer in Karel Vesely. 2011. The Kaldi speech recognition toolkit. V: *IEEE ASRU 2011 Workshop on automatic speech recognition and understanding. IEEE Signal Processing Society*.

Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li in Sanjeev Khudanpur, 2018. A Time-Restricted Self-Attention Layer for ASR. V: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), str. 5*874–5878.

RSDO. (b. d.). Dostopno na: https://www.cjvt.si/rsdo/.

Razvoj slovenščine v digitalnem okolju – RSDO: Rezultat R2.2.7: Orodje za grafemsko fonemsko pretvorbo – verzija 2, Poročilo projekta, 2022.

Christoph Schillo, Gernot A. Fink in Franz Kummert. 2000. Grapheme based speech recognition for large vocabularies. *Sixth International Conference on Spoken Language Processing*.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. V: *Seventh international conference on spoken language processing*.

Jože Toporišič. 1992. *Enciklopedija slovenskega jezika*. Cankarjeva založba. Ljubljana.

Jože Toporišič. 1991. *Slovenska slovnica*. Založba Obzorja. Maribor.

Matej Ulčar, Simon Dobrišek, Marko Robnik Šikonja. 2019. Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika*, 27 (3). Dostopno na: https://uporabna-informatika.si/index.php/ui/article/view/53 (8. 11. 2021)

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano and Kevin J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3): 328–339.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej in Tomaž Erjavec. 2013. Spoken corpus Gos 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820–4042, http://hdl.handle.net/11356/1040

Andrej Žgank, Zdravko Kačič, Frank Diehl, Jožef Juhar, Slavomir Lihan, Klara Vicsi in Gyorgy Szaszak. 2005. Graphemes as Basic Units for CrosslingualSpeech Recognition. V: *COST278 Final Workshop and ITRW on Applied Spoken Language Interaction in Distributed Environments*.

Andrej Žgank in Zdravko Kačič. 2006. Conversion from phoneme based to grapheme based acoustic models for speech recognition. *Interspeech*.