Preizkus Googlovega govornega programskega vmesnika pri samodejnem razpoznavanju govorjene slovenščine

Simon Dobrišek, David Čefarin, Vitomir Štruc, France Mihelič

Laboratorij za umetno zaznavanje, sisteme in kibernetiko,
Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{simon.dobrisek,vitomir.struc,france.mihelic}@fe.uni-lj.si, david.cefarin@gmail.com

Povzetek

Samodejni razpoznavalniki govora počasi dozorevajo v tehnologije, ki omogočajo človeku bolj naravne oblike komuniciranja z različnimi pametnimi napravami in informacijsko-komunikacijskimi sistemi. Velika svetovna podjetja, kot so Google, Microsoft, Apple, IBM in Baidu, tekmujejo pri razvoj čim bolj zanesljivih razpoznavalnikov govora, ki podpirajo čim več pomembnih svetovnih jezikov. Zaradi svoje majhnosti pa podpora govorjeni slovenščini pri govornih tehnologijah zaostaja in med velikimi svetovnimi podjetji je naš govorjeni jezik kot prvi podprl samo Google. V članku predstavljamo rezultate našega neodvisnega preizkusa Googlovega govornega programskega vmesnika pri samodejnem razpoznavanju govorjene slovenščine. Za preizkus so bile uporabljene govorne zbirke, ki jih tudi sicer uporabljamo za razvoj in preizkušanje razpoznavalnikov govorjene slovenščine.

Assessment of the Google Speech Application Programming Interface for Automatic Slovenian Speech Recognition

Automatic speech recognizers are slowly maturing into technologies that enable humans to communicate more naturally and effectively with a variety of smart devices and information-communication systems. Large global companies such as Google, Microsoft, Apple, IBM and Baidu compete in developing the most reliable speech recognizers, supporting as many of the main world languages as possible. Due to the relatively small number of speakers, the support for the Slovenian spoken language is lagging behind, and among the major global companies only Google has recently supported our spoken language. The paper presents the results of our independent assessment of the Google speech-application programming interface for automatic Slovenian speech recognition. For the experiments, we used speech databases that are otherwise used for the development and assessment of Slovenian speech recognizers.

1 Uvod

Prostoročno govorno ukazovanje avtomatiziranim in robotiziranim sistemom, govorna komunikacija z virtualnimi osebnimi asistenti na pametnih telefonih in tablicah, narekovanje diagnoz, pravnih in drugih besedil, elektronskih pisem in kratkih sporočil, samodejno tvorjenje zapisnikov sestankov, sej in telefonskih konferenc, samodejno podnaslavljanje in prepisovanje oddaj in drugih več-medijskih vsebin, preverjanje pravilnosti izgovarjave pri učenju jezika in tako dalje - vse to so primeri uporabe govornih tehnologij, ki vključujejo samodejno razpoznavanje govora.

Razvoj samodejnih razpoznavalnikov govora je že desetletja zahteven izziv, ki ga spremljajo velika pričakovanja, žal pa tudi razočaranja. Govor je človeku najbolj naraven način komuniciranja, vendar ima prav zaradi svoje naravnost določene značilnosti, ki otežujejo razvoj povsem zanesljivih samodejnih razpoznavalnikov. Zaradi okoljskih, fizioloških, sociolоških, razpoloženjskih in drugih vplivov govor posameznika izkazuje precejšnjo spremenljivost, s katero se tudi najsodobnejši razpoznavalniki govora le s težavo spopadajo. Še posebej to velja za govorjeno slovenščino, ki ima v primerjavi z nekaterimi večjimi jeziki, kot je angleščina, večje število pregibnih oblik besed in bolj prost besedni red.

Ne glede na navedeno pa je bil v zadnjih nekaj letih dosežen preboj in znaten napredek na tem področju. Razpoznavalniki govora postajajo vse bolj zanesljivi in vse več ljudi jih uporablja pri svojem vsakdanjem delu. Še posebej to velja za uporabnike, ki pri svojem delu govorno komunicirajo v enem od večjih svetovnih jezikov, kot so angleščina, španščina, portugalščina, francoščina, nemščina, italijanščina in kitajščina. Velika svetovna podjetja, kot so Google, Microsoft, Apple, IBM in Baidu, so razvila vrsto komercialnih računalniških programskih rešitev, ki vključujejo že solidno zanesljive razpoznavalnike govora, kot so Microsoft Cortana, Skype Translator, Xbox, Google Now, Apple Siri, IMB Watson Speech Recognition, Baidu Voice Search in programske rešitve podjetja Nuance.

Po črnogledih pričakovanjih pa pri navedenih programskih rešitvah podpora govorjeni slovenščini še izostaja ali vsaj zaostaja. Pozitivno presenečenje in izjema je zaenkrat le Googlov oblačni razpoznavalnik govora, ki je pred približno dvema letoma podprl tudi govorjeno slovenščino. Na demonstracijski strani Googlovega govornega programskega vmesnika[1] je že nekaj časa v spletnem brskalniku Google Chrome možno izbrati in preizkušati delovanje samodejnega razpoznavanja govorjene slovenščine. Od nedavnega je naš govorjeni jezik podprt tudi v brezplačnem Googlovem spletnem prevajalniku[2], kjer se lahko vhodno besedilo narekuje tudi v slovenščini, in tudi drugih Googlovih storitvah.

Rezultati osnovnega uporabniškega preizkusa točnosti in zanesljivosti Googlovega samodejnega razpoznavalnika govorjene slovenščine so presegli naša pričakovanja. Poleg tega pa smo dobili subjektivni vtis, da se točnost razpoznavanja sčasoma še izboljšuje. Zato smo se odločili, da Googlov govorni programski vmesnik sistematično preizkusimo z našimi obstoječimi govornimi zbirkami (Mihelič et al., 2003) in neodvisno ovrednotimo

[1] Google Web Speech API Demonstration - www.google.com/intl/en/chrome/demos/speech.html
[2] Google Translate - http://translate.google.si

zanesljivost njegovega delovanja in točnost razpoznavanja govorjene slovenščine.

V tem prispevku poročamo o naših izkušnjah in ugotovitvah, ki smo jih pridobili s preizkusom. Podajamo tudi kratek pregled napredka teh tehnologij in izpostavljamo tiste značilnosti sodobnih razpoznavalnikov govora, ki so največ pripomogli k napredku na tem področju.

2 Razvoj razpoznavalnikov govora

Prikriti Markovovi modeli (angl. Hidden Markov Models - HMM) in modeli mešanic Gaussovih porazdelitev (angl. Gaussian Mixture Models - GMM) so leta prevladovali kot najbolj uspešen zgled akustičnega in jezikovnega modeliranja, na katerem so temeljili samodejni razpoznavalniki tekočega govora z velikimi besednjaki (angl. Large-Vocabulary Continuous Speech Recognition - LVCSR). Za posebna področja uporabe z omejenim obsegom ožjega strokovnega jezika so tovrstni komercialni razpoznavalniki govora že dajali zadovoljive rezultate, denimo pri samodejnem prepisovanju narekovanih medicinskih diagnoz ali pa pravnih mnenj, strokovnih poročil in podobno.

V zadnjih nekaj letih pa je bil dosežen velik napredek z zamenjavo omenjenega nekdaj prevladujočega zgleda modeliranja z novim zgledom, ki temelji na uporabi t.i. globokih nevronskih omrežjih (angl. Deep Neural Networks - DNN) (Hinton et al., 2012; Deng et al., 2013; Siniscalchi et al., 2013; Yu in Deng, 2015). Nadaljnji znatni napredek je bil dosežen z uporabo t.i. konvolucijskih nevronskih omrežij (angl. Convolutional Neural Networks - CNN) (Sainath et al., 2013; Abdel-Hamid et al., 2014) in povratnih nevronskih omrežij (angl. Recurrent Neural Networks - RNN), ki modelirajo dolgi kratkoročni spomin (angl. Long Short-Term Memory - LSTM) (Sainath et al., 2015; Ravuri in Stolcke, 2015).

Nevronska omrežja niso novost, saj so dobro poznana že več kot pol stoletja. Za akustično modeliranje so se poskušala uporabljati že od začetka razvoja razpoznavalnikov govora (Juang in Rabiner, 2005). Vendar pa je vse do okoli leta 2010 njihov potencial na področju razpoznavanja govora ostal dokaj neizkoriščen. Zaradi omejene računske moči računalnikov in razmeroma zahtevnih učnih postopkov so bili doseženi rezultati razpoznavalnikov nezadovoljivi. Šele uspešno sodelovanje raziskovalcev z Univerze v Torontu z globalnimi podjetji, kot sta Microsof in Google, je odigralo ključno vlogo pri uveljavljanju nevronskih omrežji v komercialnih razpoznavalnikih govora (Deng 2016).

Raziskovalci so razvili učinkovita orodja za učenje globokih nevronskih omrežij, poleg tega pa je bila omogočena tudi lažja in bolj učinkovita uporaba grafičnih procesorjev za splošne računske naloge, predvsem z izdajo programske knjižnica Cuda (NVIDIA). Z novimi odkritji je bilo možno pospešiti učenje obsežnih nevronskih omrežij na velikih količinah podatkov. Obsežne govorne zbirke zvočnih posnetkov so bile na voljo, saj so raziskovalne institucije zbirale, prepisovale in označevale govorne posnetke, primerne za učenje razpoznavalnikov govora, že vsaj trideset let. Vsi ti dejavniki so omogočili uspešno sodelovanje raziskovalcev iz podjetij in univerz, kar je odprlo novo poglavje pri razvoju samodejnega razpoznavanja govora. Po letu 2010 je bilo na znanstvenih konferencah IEEE ICASSP in Interspeech sprejetih vedno več znanstvenih člankov o uporabi metod nevronskih omrežij pri razpoznavanju govora. Med komercialnimi aplikacijami pa se je uporaba nevronskih omrežij prav tako zelo hitro širila in danes jih uporablja že večina sistemov za avtomatsko razpoznavanje govora.

Pri Googlovi aplikaciji za samodejno razpoznavanje govora Google Voice so sčasoma zamenjali predkrmiljeno (angl. feedforward) omrežje s povratnim omrežjem z dolgim kratkoročnim spominom (LSTM). Ta omrežja imajo v primerjavi s prejšnjimi dodatne povratne povezave in spominske celice, ki jim omogočajo, da si »zapomnijo« pretekle podatke. Ta lastnost izboljša razpoznavanje, saj omrežje s tem modelira kontekst glasov oziroma besed, kar izboljša modeliranje govora.

2.1 Podpora za govorjeno slovenščino

Med danes najbolj razvitimi komercialnimi govornimi vmesniki je zaenkrat uspelo le Googlu podpreti tudi govorjeno slovenščino. Osnovni preizkus govornega vmesnika pri razpoznavanju govorjene slovenščine je možen na omenjeni Googlovi demonstracijski spletni strani, od nedavnega pa je v spletnem brskalniku Chrome razpoznavanje našega govorjenega jezika podprta tudi na popularni spletni strani Googlovega prevajalnika besedil. Pojavlja se tudi vse več aplikacij za pametne telefone in tablice, ki podpirajo govorjeno slovenščino, vendar po večini vse temeljijo na uporabi Googlovega vmesnika. Applov vmesnik Siri in Microsoftova Cortana pa zaenkrat še ne podpirata govorjene slovenščine in tudi ni novic, da se bo to v kratkem zgodilo.

3 Googlov govorni programski vmesnik

Googlov govorni programski vmesnik je zmogljiv, vendar razmeroma slabo dokumentiran in še vedno omejeno dosegljiv za širšo skupnost razvijalcev novih informacijsko-komunikacijskih aplikacij. Zaradi pomanjkanja dokumentacije točna sestava sistema javno še ni povsem znana, kljub temu pa so nekateri razvijalci pridobili in objavili nekaj informacij o tem, kako se vmesnik lahko uporablja.

Do Googlove govornega programskega vmesnika tako lahko dostopamo z različnimi orodji ali programskimi jeziki, ki podpirajo običajno internetno komunikacijo. Vmesnik se najbolj enostavno in najpogosteje uporablja z uporabo Javascript programskega jezika in programskega vmesnika Google Javascript Speech API. V tem primeru se aplikacije, ki uporabljajo Googlov govorni vmesnik, razvijejo kot običajne spletne aplikacije, ki se naložijo v spletni brskalnik. Omejitev je le ta, da se mora aplikacijo naložiti in uporabljati v spletnem brskalniku Google Chrome, ki zaenkrat edini podpira vse komponente govornega programskega vmesnika. Delovanje takšnih aplikacij ponazarja omenjena Googlova demonstracijska spletna stran Google Web Speech API Demonstration. Za vključevanje Googlovega govornega programskega vmesnika z uporabo Javascript API obstaja tudi dokumentacija, podprta s strani Googla (Payton, 2014).

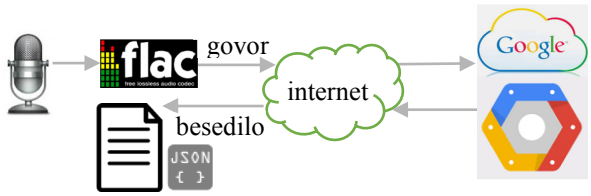
Za neposredno uporabo Googlovega govornega programskega vmesnika iz samostojne aplikacije, t.j. aplikacije, ki ni odvisna od uporabe spletnega brskalnika, pa potrebujemo ključ, ki je registriran v Googlovem oblačnem sistemu programskih vmesnikov. Ključ, ki

istoveti našo aplikacijo in ji omogoča dostop do Googlovih programskih vmesnikov, lahko pridobimo s prijavo v skupino razvijalcev aplikacij Chromium Development Group na Googlovi spletni strani https://console.developers.google.com.

Po pridobitvi ključa, ki istoveti našo aplikacijo, pa lahko Googlov govorni programski vmesnik brezplačno uporabimo le do petdesetkrat na dan in še to z omejitvijo pri dolžini obdelanih zvočnih posnetkov, ki so lahko dolgi do 15 sekund. Za intenzivnejšo uporabo govornega programskega vmesnika je potrebno plačilo po posebnem ceniku. Višina plačila je odvisna od števila uporab programskega vmesnika in dolžine vseh posnetkov, ki jih je govorni programski vmesnik obdelal.

Za raziskovalne namene in namene preizkušanja govornega programskega vmesnika pa je možna uporaba ključa, ki sicer istoveti spletni brskalnik Google Chrome. V tem primeru se lahko neka preizkusna aplikacija Googlovemu oblačnemu sistemu programskih vmesnikov dejansko predstavi na enak način kot spletni brskalnik, samo aplikacijo pa lahko spišemo v poljubnem programskem jeziku (Java, Python, C#, C++, ipd), ki podpira hkratno dvosmerno komunikacijo s strežniki z uporabo običajnih spletnih protokolov (HTTPS ipd).

Google govorni programski vmesnik lahko uporabljamo na dva načina. V prvem načinu se povežemo s Googlovim oblačnim strežnikom na internetnem naslovu https://www.google.com/speechapi/v2/recognize. Po zaključenem pošiljanju zvočne datoteke nam strežnik po istem komunikacijskem kanalu vrne rezultate razpoznavanja. Ta programski vmesnik omogoča razpoznavanje posameznih posredovanih govornih posnetkov v dolžini do okoli 15 sekund.



Slika 1: Simbolni prikaz povezave med aplikacijo in Googlovim govornim programskim vmesnikom.

Druga možnost je povezava s strežnikom na naslovih https://www.google.com/speech-api/full-duplex/v1/up in https://www.google.com/speech-api/full-duplex/v1/down. V tem primeru aplikacija s strežnikom vzpostavi dvosmerno hkratno komunikacijo v t.i. načinu »full-duplex«. To pomeni, da naša aplikacija s strežnikom hkratno komunicira v obe smeri, pri čemer strežniku pošilja zvočni signal, hkrati pa sprejema razpoznano besedilo. V tem primeru so lahko poslani govorni signali tudi daljši od 15 sekund, saj jih Googlov programski vmesnik sam po potrebi razdeli na primerne odseke in vrača rezultate razpoznavanja govora v več delih. Pri tem govorni programski vmesnik sprejema zvočne signale, kodirane v formatu FLAC (angl. Free Lossless Audio Codec) in Speex. Slednji ni standardiziran, zato je priporočena uporaba formata FLAC. Rezultate razpoznavanja strežnik vrača v datotečnem formatu JSON (angl. JavaScript Object Notation). Simbolni prikaz povezave med aplikacijo in Googlovim govornim programskim vmesnikom je prikazana na sliki 1.

Parametri internetne povezave z govornim programskim vmesnikom omogočajo različne nastavitve načina delovanja razpoznavalnika govora. Med drugim tako lahko nastavljamo stopnjo zakrivanja/filtriranja neprimernih besed (psovke ipd) v rezultatu razpoznavanja (parameter `pFilter`), izbiramo lahko število alternativnih oz. manj verjetnih rezultatov razpoznavanja, ki jih želimo pridobiti (parameter `maxAlternatives`), vključimo lahko stikalo za razpoznavanje neprekinjenega tekočega govora (parameter `continuous` - razpoznavalnik potem pošilja rezultate ob samodejno zaznanih premorih v govoru) ter stikalo, ki omogoči pošiljanje vmesnih rezultatov razpoznavanja še pred koncem izrečenih povedi (parameter `interim`) in drugi.

4 Preizkus govornega vmesnika

Preizkus Googlovega govornega vmesnika smo izvedli z uporabo lastnih izbranih preizkusnih govornih posnetkov, ki smo jih pridobili med raziskovalnim in razvojnim delom na področju govornih tehnologij. Največji del preizkusnih posnetkov smo uporabili iz naših govornih zbirk GOPOLIS (Mihelič et al., 2003) in VNTV (Žibert in Mihelič, 2000; Mihelič et al., 2003). Poleg omenjenih govornih zbirk smo za preizkus uporabili tudi nekaj dodatnih, posebej pridobljenih zvočnih posnetkov prebiranja elektronskih pisem v slovenščini (dolžine okoli 100 besed). Elektronska pisma so se prebirala na več načinov in sicer najprej počasi in razločno, nato normalno hitro in nekaj manj razločno ter nato še povsem spontano, manj razločno, z medmeti in s prekinitvami.

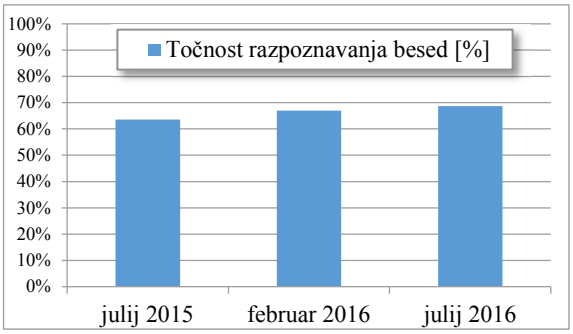
4.1 Rezultati na govorni zbirki GOPOLIS

Zbirka GOPOLIS vsebuje govorne posnetke posameznih branih povedi govornih poizvedovanj po letalskih informacijah. Po analizi dejanskih dialogov med uporabniki in uslužbenko klicnega centra za podajanje letalskih informacij je bil izdelan generator najbolj pogostih povedi. Med njimi se je nato naključno izbralo in vsakemu govorcu pripisalo od 20 do 172 povedi, ki jih je ta v bolj ali manj nadzorovanem okolju kolikor je mogoče sproščeno prebral na način, kot bi te povedi izgovoril pri komunikaciji s klicnim centrom.

Za preizkus Googlovega razpoznavalnika govora smo izbrali 10 testnih govorcev iz izvirne govorne zbirke GOPOLIS ter še 12 dodatnih govorcev, naključno izbranih med študenti, ki so imeli v okviru ene od laboratorijskih vaj pri predmetu Razpoznavanje vzorcev za nalogo zbrati neko število govornih posnetkov oz. govornih vzorcev. Pri študentskih posnetkih se je prav tako prebiralo naključno izbrane povedi, tvorjene z omenjenim generatorjem. Pri zbiranju posnetkov so bili študenti povsem prosti pri izbiri snemalnega programa, mikrofona in avdio opreme.

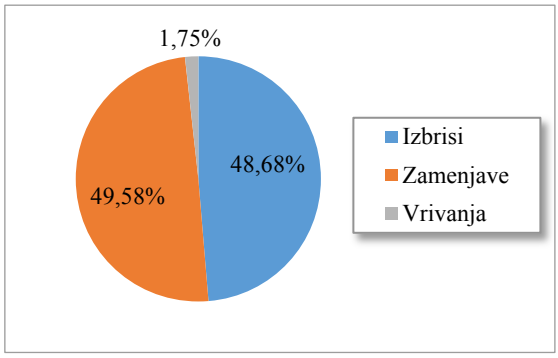
Skupaj 1925 testnih posnetkov povedi v skupnem trajanju dobre 1 ure in 37 minut se je preko posebej izdelanega programskega vmesnika, ki je temeljil na uporabi programskega orodja `curl`, poslalo Googlovemu govornemu programskemu vmesniku in pridobilo vse rezultate razpoznavanja. Pridobljene prepise govora se je nato neposredno primerjalo z referenčnimi prepisi povedi in na običajen način ugotavljalo število napak v smislu števila zamenjanih, vrinjenih in izbrisanih besed. Preizkus smo v zadnjem letu dni izvedli trikrat, ker nas je zanimalo, če se bo rezultat zaradi morebitnega prilagajanja

Googlovega razpoznavalnika govora že obdelanim govornim posnetkom v vmesnem času kaj izboljševal. Pridobljeni rezultati so potrdili, da so bila ta pričakovanja upravičena.



Slika 2: Točnost razpoznavanja besed, ugotovljena v različnih časovnih obdobjih.

Skupna točnost razpoznavanja izgovorjenih povedi se je pri preizkusih v času od julija 2015 do junija 2016 gibala od 63% do skoraj 70% pravilno razpoznanih besed. Med govorci so bile znatne razlike, saj je bila ugotovljena točnost razpoznavanja pri različnih govorcih od le 15% pa vse do 77% pravilno razpoznanih besed. Pri najslabših rezultatih je večina napak posledica dejstva, da razpoznavalnik zaradi različnih razlogov (odsotnost tišine na koncu posnetka ali prisotnost kakšnih izrazitih motenj v posnetku, kot je hrup premikanja stola ipd) ni vrnil dokončnega rezultata razpoznavanja in so se vse besede štele za izbrisane. Zaradi večjega števila lastnih imen krajev, letališč in letalskih prevoznikov, ki so omenjena v preizkusnih povedih je del napak razpoznavalnika tudi posledica napak pri njihovem ortografskem zapisu. Tako je bilo, denimo, lastno ime *Sheremetyevo* pogosto razpoznano ko *Šeremetjevo*, ali pa *Zuerich* kot *Cirih*, kar se je štelo kot napake zamenjave. Povzetek rezultatov je podan na sliki 2.



Slika 3: Pogostost različnih vrst napak pri razpoznavanju besed v preizkusnih povedih.

Analizirali smo tudi, kakšna je pogostost različnih vrst napak, torej napak izbrisa, zamenjave in izbrisa besed. Po pričakovanjih so prevladovale napake zamenjave in izbrisa, medtem ko je bil delež napak vrivanja precej manjši (slika 3). To pripisujem dejstvu, da je Googlov razpoznavalnik večkrat vrnil prazno besedilo in, kot je že bilo omenjeno, se je to navadno dogajalo, kadar posnetek povedi ni bila zaključene z dovolj premora oz. tišine.

Po pričakovanjih Googlov govorni vmesnik ni dosegel rezultatov razpoznavalnika, ki smo ga razvili sami in je bil razvit in posebej prilagojen danemu področju uporabe (poizvedovanje po letalskih informacijah). Naš razpoznavalnik je sicer že nekoliko zastarel in temelji še na uporabi prikritih Markovovih modelov (Dobrišek et al., 2006), vendar uporablja posebne kanonične akustične modele za sprotno prilagajanje govornega modela na govorne značilnosti posameznih govorcev.

Z našim razpoznavalnikom govora smo na še večjem številu podobno izbranih testnih posnetkov dosegli rezultat preko 81% pravilno razpoznavanja besed. Ključen razlog za znatno boljši rezultat pa je predvsem v tem, da naš razpoznavalnik uporablja verjetnostni jezikovni model, ki ima bistveno nižjo perpleksnost in je bistveno bolj prilagojen ožjemu področju uporabe, kot pa ga uporablja Googlov razpoznavalnik govora, saj ta temelji na splošnem generičnem jezikovnem modelu s precej višjo perpleksnostjo.

4.2 Rezultati na govorni zbirki VNTV

Poleg govorne zbirke GOPOLIS smo za preizkus Googlovega razpoznavalnika govora uporabili še 1548 televizijskih posnetkov vremenskih napovedi petih govorcev iz govorne zbirke VNTV v skupnem trajanju 1 ure in 48 minut. Rezultati in ugotovitve so podobne kot pri preizkusu s posnetki govorne zbirke GOPOLIS. Skupna točnost razpoznavanja besed je bila 62,3%, pri čemer je bil znaten deleže napak ponovno pripisan izbrisom besed pri govornih posnetkih, za katere vmesnik sploh ni vrnil prepisa, ker ti niso bili zaključeni z dovolj dolgim premorom oz. tišino. Še največje presenečenje je bilo, da je bil rezultati (81,1 %) pri ženski govorki (napovedovalka Tanja Cegnar) v povprečju znatno boljši od rezultatov moških govorcev (od 54% do 70%). Navadno se namreč pri preizkusih samodejnih razpoznavalnikov govora izkaže ravno obratno in imajo ti večje težave z ženskimi govorkami.

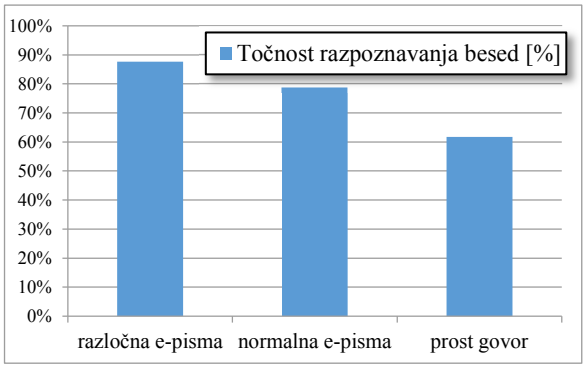
4.3 Rezultati pri narekovanju e-pisem

Zadnji preizkus, ki smo ga izvedli v okviru te raziskave, pa se je nanašal na razvoj uporabne aplikacije govornega narekovanja elektronskih pisem oz. sporočil. Za preizkus smo pridobili posnetke enega moškega govorca, ki je prebiral elektronska sporočila v skupnem obsegu približno 100 besed in imajo obliko pritožbe ali povabila. Besedila imajo na začetku in na koncu preprost pozdrav, sama vsebina pa je sestavljena iz stavkov in fraz, ki so pogosta pri poljudnem in prostem izražanju. Elektronska pisma namreč navadno ne vsebujejo zapletenih strokovnih izrazov, preizkus pa je bil namenjen ugotavljanju primernosti razpoznavalnika za uporabo pri narekovanju vsakdanjih preprostih nestrokovnih besedil.

Govorec je elektronska sporočila izgovarjal na dva načina. V prvem načinu je bilo izgovarjanje počasno in zelo razločno. V drugem primeru pa so bili isti primeri sporočil prebrani normalno hitro in nekaj manj razločno.

Pri tretjem preizkusu smo uporabili še tri posnetke prostega govora, kjer ni bilo branja besedila, ampak si je govorec besedilo sproti izmišljeval kar med samim govorom. Tak način govora je bil neenakomeren in je imel precej prekinitev in tudi nekaj govornih mašil oziroma medmetov. Vsebina govora se je nanašala večinoma na splošen opis dneva, kot so razmere na cesti ali izvedena opravila. Besedila so bila dolga okoli 150 besed in prav tako kot pri branju elektronskih pisem niso vsebovala

strokovnih ali žargonskih besed. Rezultat preizkusa je podan na Sliki 4.



Slika 4: Ugotovljena točnost razpoznavanja treh dodatnih vrst besedil.

Iz rezultatov je razvidno, da malo počasnejša in razločna izgovarjava znatno pripomore k višji točnosti razpoznavanja in da v tem primeru postane Googlov razpoznavalnik govora že uporaben za marsikatero aplikacijo

5 Zaključek

V članku so podani rezultati izvedenih preizkusov Googlovega razpoznavalnika pri razpoznavanju govorjene slovenščine. Rezultati se odvisno od govorca in podatkovne zbirke gibljejo nekje od 40 do 12 odstotne ocenjene napake razpoznavanja besed. Ta rezultat je nekaj slabši od rezultata približno 8 odstotkov napačno razpoznanih besed, kot ga je za govorjeno angleščino objavil sam Google, vendar zaradi zelo obsežnega splošnega jezikovnega modela in zahtevnih posebnosti slovenščine, ki otežuje zanesljivost samodejnega razpoznavanja, rezultat presega pričakovanja. Poleg tega pa se je izkazalo, da se rezultat sčasoma izboljšuje, saj je bil ugotovljen napredek pri preizkusih na istih posnetkih, ki so bili izvedeni v različnih časovnih obdobjih. Ugotovljeni napredek pripisujemo dejstvu, da se akustične posnetke, ki se pošiljajo v obdelavo programskemu vmesniku v Googlovem računalniškem oblaku, zelo verjetno Google zbira in uporablja za prilagajanje in izboljševanje njegovih akustičnih in jezikovnih modelov.

6 Literatura

Ossama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn in Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions On Audio, Speech, And Language Processing,* 22(10): 1533-1545.

Li Deng, Geoffrey Hinton, Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: An overview. V: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP-2013*, str. 8599-8603, Vancouver, BC, Kanada. IEEE – Institute of Electrical and Electronics Engineers.

Li Deng. 2016. Industrial Technology Advances: Deep learning --- from speech recognition to language and multimodal processing. V: *APSIPA Transactions on Signal and Information Processing*, Cambridge University Press.

Simon Dobrišek, Boštjan Vesnicer, Jerneja Žganec Gros in France Mihelič. 2006. Uporaba kanoničnega govornega akustičnega modela za prilagajanje prostora govornih akustičnih značilk. V: *Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.). Jezikovne tehnologije : zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, str. 89-92, Ljubljana, Slovenija. Institut Jožef Stefan.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath in Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Biing-Hwang Juang, Lawrence Rabiner. 2005. *Automatic speech recognition–a brief history of the technology development,* Georgia Institute of Technology. Atlanta Rutgers University and the University of California, Santa Barbara.

France Mihelič, Jerneja Žganec Gros, Simon Dobrišek, Janez Žibert, Nikola Pavešić. 2003. Spoken language resources at LUKS of the University of Ljubljana. *International Journal of Speech Technology*, 6(3): 221-232.

Travis Payton. 2014. *Google Speech Api Information and Guidelines*. Google.

Suman Ravuri in Andreas Stolcke. 2015. Recurrent Neural Network and LSTM Models for Lexical Utterance Classification. V: *Proceedings of the Annual Conference of the International Speech Communication Association – INTERSPEECH 2015*, str. 135-139, Dresden, Germany. ISCA - International Speech Communication Association.

Tara N. Sainath, Abdel-rahman Mohamed, Brian Kingsbury in Bhuvana Ramabhadran. 2013. Deep Convolutional Neural Networks for LVCSR. V: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP-2013*, str. 8614-8618, Vancouver, BC, Kanada. IEEE – Institute of Electrical and Electronics Engineers.

Tara N. Sainath, Oriol Vinyals, Andrew Senior in Hasim Sak. 2015. Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks. V: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2015)*, str. 4580-4584, Brisbane, Australija. IEEE – Institute of Electrical and Electronics Engineers.

Sabato Marco Siniscalchi, Dong Yu, Li Deng in Chin-Hui Lee. 2013. Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing*, 106 (2013): 148–157.

Sabato Marco Siniscalchi, Dong Yu, Li Deng in Chin-Hui Lee. 2013. Speech Recognition Using Long-Span Temporal Patterns in a Deep Network Model, *IEEE Signal Processing Letters*, 20 (3): 201-204.

Dong Yu in Li Deng. 2015. *Automatic Speech Recognition: A Deep Learning Approach*. Springer-Verlag, London.

Janez Žibert in France Mihelič. 2000. Govorna zbirka vremenskih napovedi. V: *Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.). Jezikovne tehnologije : zbornik konference*, str. 108-111, Ljubljana, Slovenija. Institut Jožef Stefan.