Govorne tehnologije u programskom jeziku Python za hrvatski jezik

Sirovec, Lucia

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:369145

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2022./2023.

Lucia Sirovec

Govorne tehnologije u programskom jeziku Python za hrvatski jezik

Diplomski rad

Mentor: dr. sc. Ivan Dunđer, docent

Zagreb, srpanj 2023.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Za ekipu ispred knjižnice koja nikad nije prestala biti moj oslonac i za mog Domagoja koji mi je u svemu nesebična podrška.

Sadržaj

1. Uvod

U današnjem digitalnom dobu, sposobnost učinkovitog pretvaranja govornog jezika u pisani tekst postaje sve važnija. Područje tehnologije pretvorbe govora u tekst, poznato i kao automatsko prepoznavanje govora (engl. *automatic speech recognition*, ASR), posljednjih je godina doživjelo značajan napredak. Ovaj napredak otvorio je nove mogućnosti i izazove u raznim domenama koje uključuju i pristupačnost (engl. *accessibility*), komunikaciju, transkripciju i analizu podataka. Ovaj rad istražit će trenutno stanje tehnologije pretvorbe govora u tekst, njezine temeljne tehnike i prikazati njezine primjene u različitim industrijama.

Proces pretvaranja govornog jezika u pisani tekst tradicionalno je bio naporan i dugotrajan zadatak. Međutim, s pojavom strojnog učenja i tehnika obrade prirodnog jezika, automatsko prepoznavanje govora pokazalo se kao rješenje koje bi moglo skratiti proces i učiniti ga učinkovitim. Koristeći sofisticirane algoritme i osobito arhitekturu neuronskih mreža, tehnologija pretvorbe govora u tekst postavlja dobre temelje za premošćivanje jaza između govornog i pisanog jezika, omogućujući besprijekoran i točan prijepis te mogućnost daljnje obrade zapisa.

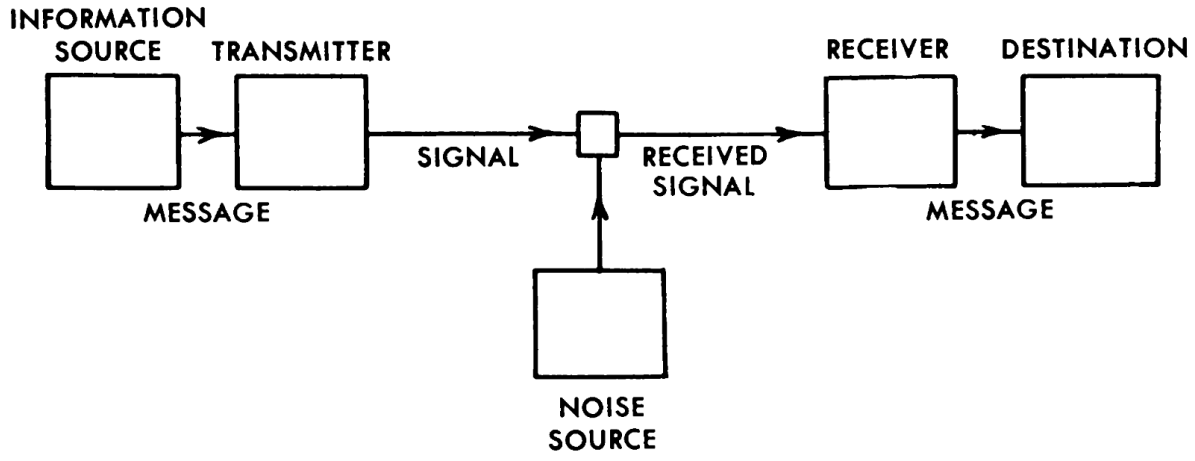
Ovaj rad ima za cilj pružiti sveobuhvatan pregled tehnologije pretvorbe govora u tekst, pokrivajući njezine temeljne tehnike, izazove i primjene. Rad će se baviti temeljnim komponentama tehnologije govora u tekst, počevši od govornog signala i stvaranja govora te kvalitetama govornog signala. Zatim će se istražiti tehnike obrade signala i izdvajanja značajki koje se koriste za transformaciju govornog zvučnog signala u smislene prikaze koji se mogu obraditi modelima strojnog učenja. Bit će prikazani različiti pristupi prepoznavanju govora i pretvorbi govora u tekst, uključujući skrivene Markovljeve modele (engl. *hidden Markov model*, HMM), duboke neuronske mreže (engl. *deep neural network*, DNN) kao i hibridne modele koji kombiniraju prednosti oba.

Ovaj će rad također prikazati trenutno stanje tehnologije pretvorbe govora u tekst i prepoznavanja govora za hrvatski jezik. Također će biti navedeni i neki alati kojima je to moguće postići. Nadalje, bit će prikazano testiranje i rezultati testiranja pretvorbe govora u tekst uz pomoć nekih od navedenih alata koje je moguće koristiti uz pomoć programskog jezika Python.

2. Teorija informacije i komunikacije

Matematička teorija informacije i komunikacije Claudea Shannona i Warrena Weavera stvorena je 1964. godine. Komunikacija je pojam koji su u svom radu koristili u njegovom širem smislu kao svaki postupak kojim jedan um utječe na drugi. To uključuje govor, pismo, ali i glazbu, umjetnost i sva ljudska ponašanja. Shannon i Weaver na početku svog rada uvode i tri problema komunikacije: tehnički problem, semantički problem i problem učinkovitosti. Tehnički se problem odnosi na točnost prijenosa podataka od pošiljatelja do primatelja preko medija kojim se prenosi poruka (zapisani jezik, telefonski ili radio prijenos glasa, televizijski signal). Semantički problem odnosi se na interpretaciju značenja prenesene poruke od strane primatelja, u usporedbi sa značenjem koje je pošiljatelj htio poslati. Problem učinkovitosti odnosi se na željeni učinak prenesene poruke. Sva tri problema komunikacije međusobno su isprepletena: bez primjereno riješenog tehničkog problema u nekoj komunikaciji – nema komunikacije, a semantički problem i problem učinkovitosti u određenim slučajevima su upravo i tehnički problemi (Weaver & Shannon, 1964).

Shannon i Weaver stvorili su grafički prikaz komunikacijskog sustava, tj. komunikacijskog kanala (slika 1). Prikaz je općenit te ga je moguće primijeniti na razne vrste komunikacije kroz razne medije (Weaver & Shannon, 1964).



Slika 1 - grafički prikaz komunikacijskog sustava (Weaver & Shannon, 1964)

Informacijski izvor (engl. *information source*) stvara poruku te ju putem odašiljača (engl. *transmitter*) odašilje u komunikacijski kanal prema primatelju poruke. Odašiljač poruku pretvara u odgovarajući signal za određeni kanal, što može biti pisana poruka, telefonski ili televizijski signal, slika, glazba i tako dalje. S primateljeve strane (engl. *destination*) kanala prijemnik (engl. *receiver*) dekodira primljeni signal natrag u poruku. U procesu prijenosa

signala od pošiljatelja prema primatelju redovito se dodaju i drugi nepoželjni signali, tj. stvara se buka (engl. *noise*) u komunikacijskom kanalu koja može biti bilo kakva distorzija zvuka, slike ili samo greška u prijenosu (Weaver & Shannon, 1964).

Govor, odnosno prepoznavanje govora i pretvorba govora u tekst kao svoju glavnu svrhu imaju komunikaciju (Rabiner & Schafer, 2007). Prepoznavanje govora i pretvorba govora u tekst korisne su za poboljšavanje komunikacije između ljudi, ali i ljudi i računala. Kao i mnogo drugih tehnologija, tehnologija prepoznavanja govora temelji se upravo na teoriji informacije i komunikacije Shannona i Weavera, tj. na rješavanju problema predstavljenog komunikacijskog kanala (Yu & Deng, 2015).

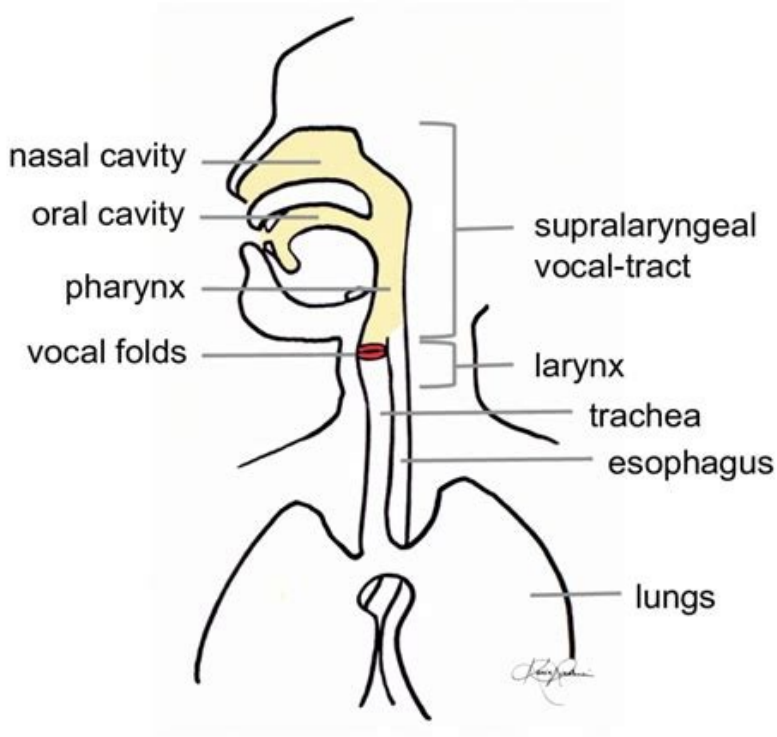
3. Govor

Govor je zvučno sredstvo ostvarenja jezika, tj. sustav verbalnih i neverbalnih znakova koji imaju značenje i koji se koriste u komunikaciji (Hrvatska enciklopedija, 2021). Prema Shannonovoj teoriji informacije, svaku je poruku moguće opisati nizom simbola koje je potom moguće i kvantificirati po količini informacije koja se nalazi u dijelu poruke (Weaver & Shannon, 1964). U mnogim elektroničkim komunikacijskim sustavima, informacija koja se prenosi, kodirana je u obliku kontinuirano promjenjivog valnog oblika koji se može prenositi, snimati, njime se može manipulirati, i na kraju ga se može dekodirati. U slučaju govora, temeljni analogni oblik poruke je akustički valni oblik, kojeg nazivamo govornim signalom. Govorni signal može se pretvoriti u električni val pomoću mikrofona. Njime se dalje upravlja obradom analognog i digitalnog signala, a na kraju je moguće signal pretvoriti natrag u njegov akustični oblik preko zvučnika ili slušalica (Rabiner & Schafer, 2007).

3.1. Vokalni organi i kreiranje govora

Zvuk nastaje brzim kretanjem zraka. Ljudi proizvode većinu zvukova u govornim jezicima izbacivanjem zraka iz pluća kroz dušnik (engl. *trachea*) i zatim kroz usta ili nos. Dok prolazi kroz dušnik, zrak obično prolazi kroz grkljan (engl. *larynx*) koji je poznat kao Adamova jabučica ili glasovna kutija gdje se nalaze i dva nabora mišića koje nazivamo glasnicama (engl. *vocal folds*). Područje iznad dušnika naziva se vokalnim traktom (engl. *vocal tract*). Vokalni trakt se sastoji od oralnog trakta (engl. *oral cavity*) i nosnog trakta (engl. *nasal cavity*). Nakon što zrak napusti dušnik, može izaći iz tijela kroz usta ili nos. Većina zvukova nastaje zrakom koji prolazi kroz usta. Zvukovi kao najmanje jezične jedinice koje nemaju značenja nazivaju se fonemi. Fonemi se dijele prema tome nastaju li prolaskom zraka kroz nos ili usta, te koriste li se pri stvaranju fonema (i na koji način) usne te jezik te se promatra položaj jezika u usnoj

šupljini (Jurafsky & Martin, 2019). Organi potrebni za proizvodnju zvukova prikazani su na slici 2.



Slika 2 - organi potrebni za govor (Pisanski, 2014)

Fonemi se dijele u dvije glavne klase: suglasnike i samoglasnike. Obje vrste zvukova nastaju kretanjem zraka kroz usta, grlo ili nos. Suglasnici nastaju ograničavanjem ili blokiranjem protoka zraka na neki način, a mogu biti zvučni ili nezvučni. Samoglasnici imaju manje prepreka, obično su zvučni i općenito su glasniji i dugotrajniji od suglasnika (Jurafsky & Martin, 2019).

Suglasnici se međusobno razlikuju po načinu na koji je napravljeno ograničenje protoka zraka, na primjer, potpunim zastojem zraka ili djelomičnim zastojem protoka. Ova značajka se zove način artikulacije suglasnika, odnosno način tvorbe suglasnika. Kombinacija mjesta i načina artikulacije je obično dovoljna za jedinstvenu identifikaciju suglasnika (Jurafsky & Martin, 2019). U hrvatskom jeziku se razlikuju šumnici (opstruenti) koji imaju veću zapreku i zvonačnici (sonanti) koji imaju manju zapreku. Samoglasnici se međusobno razlikuju s obzirom na položaj jezika u usnoj šupljini (Hrvatska enciklopedija, 2021).

3.2. Fonetski prikaz govora

Govor se fonetski može predstaviti konačnim skupom simbola, odnosno najmanjim zvučnim jedinicama bez značenja koje nazivamo fonemima. Sam prikaz fonema vuče svoje korijene iz početaka ljudske pismenosti. Najraniji dokazi pismenosti bili su logografski oblici pisma gdje

je jedan simbol označavao riječ na koju se odnosi. No, neki su simboli označavali zvuk koji je bio samo dio riječi, upravo ono što danas nazivamo fonemom. Ideja sustava pisanja temeljenog na zvuku, tj. da je izgovorena riječ sastavljena od manjih govornih jedinica je osnova modernih algoritama za prepoznavanje i sintezu govora. Broj fonema jezika ovisi o samom jeziku i profinjenosti analize jezika. Za većinu jezika broj fonema je između 32 i 64 (Jurafsky & Martin, 2019).

Međunarodna fonetska abeceda (engl. *international phonetic alphabet*, IPA) je standard u stalnom razvoju kojeg je izvorno razvila Međunarodna fonetska udruga (engl. *International Phonetic Association*) 1888. godine s ciljem da se zapišu i transkribiraju svi zvukovi svih ljudskih jezika (Jurafsky & Martin, 2019). Na slici 3 prikazani su načini tvorbe glasova i njihov prikaz međunarodnom fonetskom abecedom.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

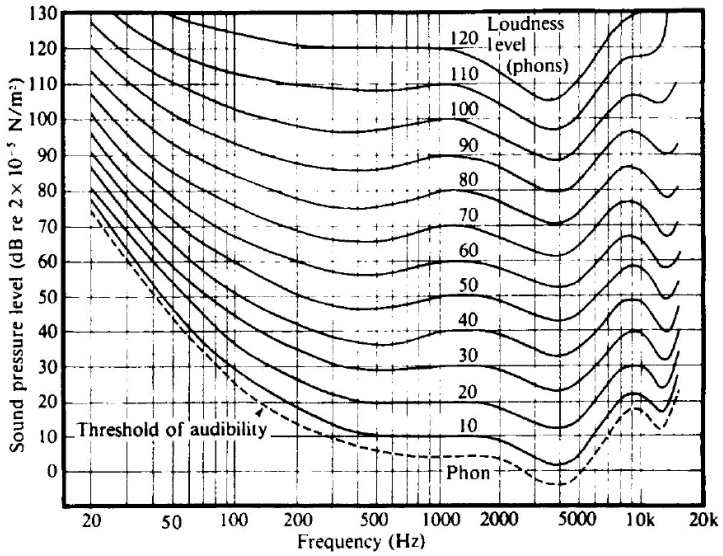
Slika 3 - međunarodna fonetska abeceda (InternationalPhoneticAlphabet.org, 2015)

3.3. Kvalitete govora

Govor, kao i zvuk ima važna svojstva na temelju kojih se razlikuje od pozadinske buke i olakšava prepoznavanje govora. Dva važna svojstva vezana uz frekvenciju govora i njegov intenzitet su visina tona (engl. *pitch*) i glasnoća (engl. *loudness*) (Jurafsky & Martin, 2019).

Visina tona zvuka i govora je mentalni osjet ili perceptivni korelat osnovne frekvencije tog zvuka. Općenito, ako zvuk ima višu osnovnu frekvenciju mi ju doživljavamo kao da zvuk ima viši ton. Odnos više frekvencije i mentalnog osjeta nije linearan jer ljudski sluh ima različitu oštrinu na različite frekvencije. Ljudska percepcija visine tona je najtočnija između 100 Hz i 1000 Hz te u tom rasponu visine mentalni osjet linearno korelira s frekvencijom. Ljudski sluh frekvencije iznad 1000 Hz doživljava manje točno te u tom rasponu visina tona korelira logaritamski, za razliku od nižih frekvencija (Jurafsky & Martin, 2019).

Glasnoća zvuka je mentalna percepcija snage zvuka i govora. Zvukovi većih amplituda percipiraju se kao glasniji, no kao i kod visine tona zvuka, odnos nije linearan. Ljudi imaju veću mogućnost razlučivosti kod manjih glasnoća zvuka te je uho osjetljivije na male razlike u amplitudi. Osim toga, postoji složen odnos između snage zvuka i govora, frekvencije i percipirane glasnoće u smislu da se određeni frekvencijski rasponi percipiraju glasnijima bez obzira na svoju snagu (Jurafsky & Martin, 2019). Glasnoća se kvantificira odnosom stvarne razine zvučnog tlaka čistog tona koja je izražena u decibelima (dB) u odnosu na standardnu referentnu razinu s percipiranom glasnoćom istog tona (u mjernoj jedinici fon) u rasponu ljudskog sluha (koji je između 20 Hz i 20000 Hz). Slika 4 prikazuje taj odnos percepcije glasnoće na temelju zvučnog tlaka i frekvencije zvuka (Rabiner & Schafer, 2007).



Slika 4 - razine glasnoće koje percipira ljudsko uho (Rabiner & Schafer, 2007)

4. Prepoznavanje govora

Automatsko prepoznavanje govora (engl. *automatic speech recognition*, ASR) je aktivno područje istraživanja više od pet desetljeća. Uvijek se smatralo važnim mostom u njegovanju bolje komunikacije čovjek-čovjek i čovjek-stroj. U prošlosti, za razliku od danas, govor nije bio važan faktor u komunikaciji čovjeka i stroja. Dijelom to možemo pripisati tome što tehnologija u to vrijeme nije bila dovoljno dobra da prijeđe prag korisnosti za većinu korisnika u stvarnom svijetu u većini stvarnih uvjeta korištenja, a dijelom zato što su u mnogim situacijama alternativni načini komunikacije poput tipkovnice i miša značajno nadmašivali govor u komunikacijskoj učinkovitosti, njenim ograničenjima i na kraju, točnosti (Yu & Deng, 2015).

4.1. Razvoj tehnologije

Posljednjih godina govorna tehnologija počela je mijenjati način na koji živimo i radimo te je postala jedan od primarnih načina interakcije ljudi s uređajima. Ovaj trend započeo je zahvaljujući napretku postignutom u nekoliko ključnih područja. Prvenstveno, opažanje da se svake dvije godine udvostručuje broj tranzistora u gustom integriranom krugu, tj. Mooreov zakon još uvijek djeluje. Računalna snaga dostupna danas kroz višejezgrene procesore, grafičke procesorske jedinice opće namjene (engl. *general purpose graphical processing units*, GPGPU) i klastere kompjuterskih i grafičkih procesorskih jedinica (engl. *computer processing unit/graphical processing unit*, CPU/GPU) nekoliko je redova veličine više nego što je bilo dostupno par godina prije. Eksponencijalni skok u mogućnostima procesiranja čini mogućim i lakšim treniranje moćnijih i složenijih modela na području strojnog učenja. Računalno zahtjevniji modeli značajno su smanjili stopu pogrešaka unutar sustava za prepoznavanje govora. Nadalje, sada možemo pristupiti puno većoj količini podataka nego prije, zahvaljujući stalnom rastu interneta i računarstva u oblaku. Izgradnjom modela velikih podataka koji su prikupljeni iz stvarnih situacija upotrebe možemo eliminirati mnoge pretpostavke modela napravljene prije te sva nova znanja čine sustave robusnijim. Također, mobilni uređaji, nosivi uređaji poput pametnih satova, inteligentni I-o-T (engl. *Internet of Things*) uređaji i infotainment (engl. *information-entertainment*) medijski sustavi u vozilima postali su popularni. Na ovim uređajima i sustavima, alternativna interakcija umjesto govora drugim sredstvima kao što su tipkovnica i miš manje su prikladni od onih u osobnim računala. Govor, koji je prirodni način komunikacije između ljudi, i a vještina koju većina ljudi već ima, predstavlja se kao najjednostavniji i najsmisleniji način komunikacije s takvim uređajima i sustavima (Yu & Deng, 2015).

4.2. Primjena govorne tehnologije prepoznavanja govora

Postoje mnoge primjene u kojima govorna tehnologija igra važnu ulogu. Ove se aplikacije mogu klasificirati kao aplikacije koje mogu pomoći u poboljšanju komunikacije čovjek-čovjek (engl. *human-human communication*, HHC) ili u poboljšanju komunikacije čovjek-stroj (engl. *human-machine communication*, HMC) (Yu & Deng, 2015).

Govorna tehnologija može ukloniti prepreke kod interakcija između ljudi. Ljudi koji govore različitim jezicima u prošlosti su trebali ljudskog tumača da bi mogli razgovarati jedan s drugim. Ovo postavlja značajno ograničenje na to s kim ljudi mogu komunicirati i kada se komunikacija može dogoditi. Na primjer, odlazak na putovanje u zemlju s potpuno nepoznatim jezikom može predstavljati izazov potencijalnom putniku. Ova se prepreka može ublažiti

sustavima za prevođenje govora u govor (engl. *speech to speech*, S2S), čiji je prvi dio upravo prepoznavanje govora. Osim što ga mogu koristiti putnici, prevoditeljski sustavi govora u govor također se mogu integrirati u komunikacijske alate kao što su Zoom ili Skype te omogućiti ljudima koji govore različite jezike da slobodno međusobno komuniciraju na daljinu. Govorna tehnologija također može pomoći komunikaciji između ljudi na druge načine. Na primjer, u jedinstvenom sustavu za razmjenu poruka, podsustav za transkripciju govora može se koristiti za pretvaranje glasovne poruke koje je ostavio pošiljatelj u tekst. Transkribirani tekst tada se može lako poslati primatelju putem e-pošte, izravnih poruka ili kratkih poruka. U drugom primjeru, tehnologija prepoznavanja govora može se koristiti za diktiranje kratkih poruka kako bi se smanjio napor potreban korisnicima za slanje kratkih poruka. Tehnologija prepoznavanja govora također se može koristiti za prepoznavanje govora u mnogim drugim situacijama (primjerice predavanja, rasprave, politički govori) te je potom moguće indeksirati govor za snalaženje u izrečenom govoru i pronalazak informacija koje su korisnicima zanimljive. Na kraju, tehnologija pretvorbe govora u tekst može značajno poboljšati kvalitetu života osobama sa slušnim poteškoćama (Yu & Deng, 2015).

Govorne tehnologije također mogu uvelike poboljšati komunikaciju između čovjeka i stroja. Najpopularnije primjene u ovoj kategoriji uključuju glasovno pretraživanje, osobnog digitalnog asistenta, igre, interakcijske sustave pametnih domova i infotainment medijske sustave u vozilu. Aplikacije glasovnog pretraživanja omogućuju korisnicima traženje informacija na raznim tražilicama i bazama podataka izravno putem govora. Značajno smanjuju napor potreban korisnicima za unos upita za pretraživanje. Ovih dana, aplikacije za glasovno pretraživanje vrlo su popularne u mobilnim uređajima. Osobni digitalni asistenti (engl. *personal digital assistant*, PDA) izrađuju se već desetljećima. Međutim, postali su popularni relativno nedavno nakon što je Apple pustio Siri sustav u iPhone mobilnim uređajima. Od tada su mnoge druge tvrtke plasirale slične proizvode. Osobni digitalni asistent „zna“ informacije pohranjene u mobilnom uređaju, neka svjetska znanja i povijest korisničkih interakcija sa sustavom, te integracijom svih tih znanja mogu bolje služiti korisnicima. Osobni digitalni asistenti sve zadatke koje obavljaju (primjerice biranje telefonskog broja, zakazivanje sastanka i ažuriranje kalendara, traženje odgovora, puštanje glazbe, glasovno diktiranje za stvaranje pisama, bilješki i drugih dokumenata) čine nakon što korisnik da glasovnu naredbu. Prepoznavanje govora izgovorenog imena u mobitelima omogućuje glasovno biranje koje može automatski birati broj povezan s prepoznatim imenom. Imena iz imenika s više od nekoliko stotina imena mogu se lako prepoznati i biraju se pomoću jednostavne tehnologije prepoznavanja govora (Rabiner &

Schafer, 2007). Iskustvo igranja može se znatno poboljšati ako su igre integrirane s govornim tehnologijama. Na primjer, u nekim Microsoftovim igrama za Xbox igrači mogu razgovarati s likovima i tražiti informacije i davati naredbe. Interakcijski sustavi pametnih domova i infotainment medijski sustavi u vozilima vrlo su slični kada govorimo o funkcionalnostima. Ovi sustavi omogućuju korisnicima interakciju s njima putem govora kako bi korisnici mogli puštati glazbu, tražiti informacije ili kontrolirati sustav (Yu & Deng, 2015).

5. Obrada govornog signala

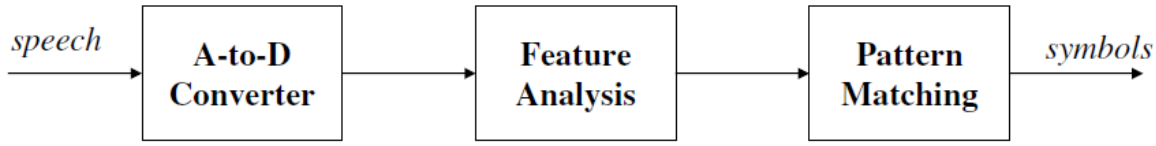
Obrada govornog jezika odnosi se na tehnologije povezane s prepoznavanjem govora, pretvaranjem teksta u govor i razumijevanje govornog jezika. Sustav govornog jezika ima barem jedan od sljedeća tri podsustava: sustav za prepoznavanje govora koji pretvara govor u riječi, sustav pretvaranja teksta u govor koji prenosi govorne informacije i razumijevanje govornog jezika te sustav koji preslikava riječi u djela i koji planira radnje koje pokreće sustav (Huang, Acero, & Hon, 2001). Postoji značajno preklapanje u temeljnim tehnologijama za ova tri potpodručja, no u nastavku rada fokus će biti na prvom, tj. sustavu za prepoznavanje govora koji pretvara govor u riječi, odnosno u tekst (Yu & Deng, 2015).

Od svojih početaka, sustavi govornog jezika prolazili su kroz razvojne stadije tehnologije usporedno s razvojem svih potrebnih komponenata za modeliranje cjelovitog sustava. Moguće je definirati ručno stvorena pravila koja su razvijena za sustave govornog jezika, no ona imaju ograničen uspjeh zbog teškog sastavljanja skupa pravila potrebnih za kvalitetno prepoznavanje govora. Posljednjih desetljeća puno se više primjenjuju statistički pristupi temeljeni na podacima te su oni postigli rezultate koji se obično temelje na modeliranju govornog signala korištenjem dobro definirane statistike i na algoritmima koji mogu automatski izvući znanje iz podataka. Pristup vođen podacima može se u osnovi promatrati kao problem prepoznavanja uzoraka. Zapravo, sve tehnologije sustava govornog jezika (prepoznavanje govora, pretvaranje teksta u govor i razumijevanje govornog jezika) mogu se smatrati problemima prepoznavanja uzoraka. Uzorci se mogu prepoznavati tijekom rada programa sustava ili su identificirani tijekom izgradnje sustava što omogućava oblikovanje osnova generativnih modela koji se stvaraju za vrijeme rada sustava. Primjer toga su prozodijski predlošci potrebni za sintezu teksta u govor. Dok se uglavnom koristi i zagovara statistički pristup, nipošto se ne isključuje pristup inženjeringa znanja i stvaranje pravila dobivenih iz razmatranja problema. Ako postoji dobar skup pravila u određenom problemskom području, uopće nema potrebe koristiti statistički

pristup. Stoga je pristupe temeljene na pravilima i statističke pristupe najbolje promatrati kao komplementarne (Huang, Acero, & Hon, 2001).

5.1. Problemi prepoznavanja govora

Velik dio digitalne obrade govora brine o automatskom izdvajanju informacija iz govornog signala. Većina sustava uključuje neku vrstu podudaranja uzoraka. Slika 5 prikazuje dijagram generičkog pristupa problemu sparivanja uzoraka u digitalnoj obradi govornog signala. Problemi na koje se nailazi prilikom obrade govornog signala su sljedeći: prepoznavanje govora, gdje je cilj izdvojiti poruku iz govornog signala; prepoznavanje govornika, gdje je cilj prepoznati tko govori; verifikacija govornika, gdje je cilj provjera govornikovog identiteta iz analize njegovog govornog signala; uočavanje riječi (engl. *word spotting*), što uključuje praćenje govornog signala za pojavljivanje određenih riječi ili fraza; i automatsko indeksiranje govornih zapisa temeljenih na prepoznavanju (ili uočavanju) izgovorenih ključnih riječi (Rabiner & Schafer, 2007).

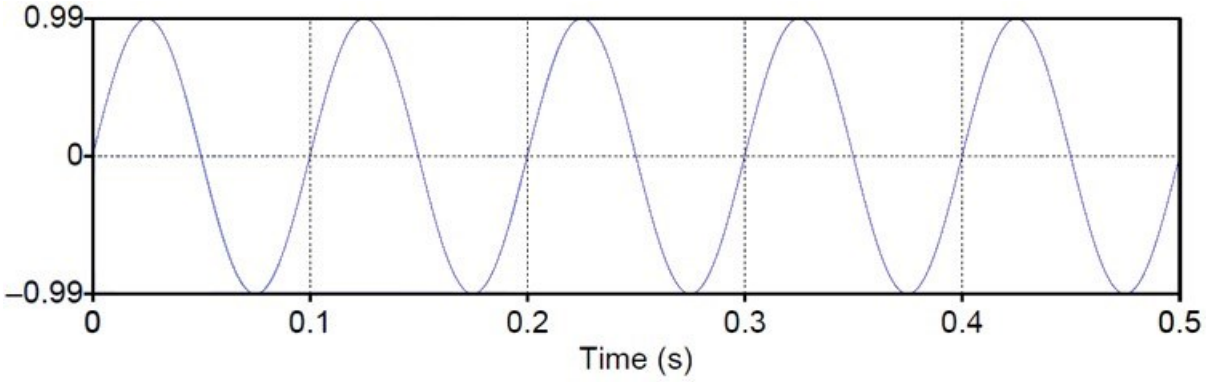


Slika 5 - dijagram sustava obrade govornih signala (Rabiner & Schafer, 2007)

Prvi blok u sustavu usklađivanja uzoraka pretvara govorni signal analognog oblika u digitalni oblik pomoću pretvarača analognog u digitalni signal (engl. *analogue-to-digital converter*, A-to-D). Modul za analizu značajki (engl. *feature analysis*) pretvara uzorkovani govorni signal u skup vektora obilježja. Često se iste tehnike analize u kodiranju govora također koriste za izvođenje vektora značajki. Konačni blok u sustavu, tj. blok za usklađivanje uzorka (engl. *pattern matching*) dinamički i vremenski usklađuje skup vektora značajki koji predstavljaju govorni signal sa spojenim skupom pohranjenih uzoraka, i odabire pridruženi identitet s uzorkom koji se najviše podudara s vremenski usklađenim skupom vektora obilježja govornog signala. Simbolički izlaz sustava obrade govornih signala može se sastojati od skupa prepoznatih riječi u slučaju prepoznavanja govora, u slučaju prepoznavanja govornika može biti prepoznavanje govornika, ili u slučaju verifikacije govornika može biti odluka sustava hoće li prihvatiti ili odbiti tvrdnju o identitetu govornika (Rabiner & Schafer, 2007).

5.2. Valni oblik govora

Akustička analiza govora temelji se na sinusnoj i kosinusnoj funkciji. Važne karakteristike valova su amplituda i frekvencija. Amplituda je maksimalna vrijednost na osi y, a frekvencija je broj ciklusa po sekundi, ujedno ih nazivamo Hertzima, skraćeno Hz. Slika 6 prikazuje jedan sinusni val, odnosno funkciju y= A**sin*(2π*ft*), gdje je amplituda A=1, a frekvencija *f* deset Hertza (Jurafsky & Martin, 2019).



Slika 6 - sinusni val frekvencije 10 Hz i amplitude 1

Unos govornog signala u sustav za prepoznavanje govora, poput slušanja ljudskim uhom je složen niz promjena tlaka zraka. Te promjene tlaka zraka potječu od govornika, a uzrokovane su specifičnim načinom na koji zrak prolazi kroz dušnik i izlazi iz usne ili nosne šupljine. Govorni signal prikazujemo zvučnim valovima gdje su vidljive promjene tlaka zraka tijekom vremena. Pretvaranje govornog signala u valni oblik može se opisati kao postavljanje zamišljene okomite ploče koja blokira valove tlaka zraka (primjerice u mikrofonu ispred usta govornika ili bubnjić u uhu slušatelja). Valni oblik govora prikazuje količinu kompresije ili razrjeđivanje (nekompresija) molekula zraka na zamišljenoj ploči (Jurafsky & Martin, 2019).

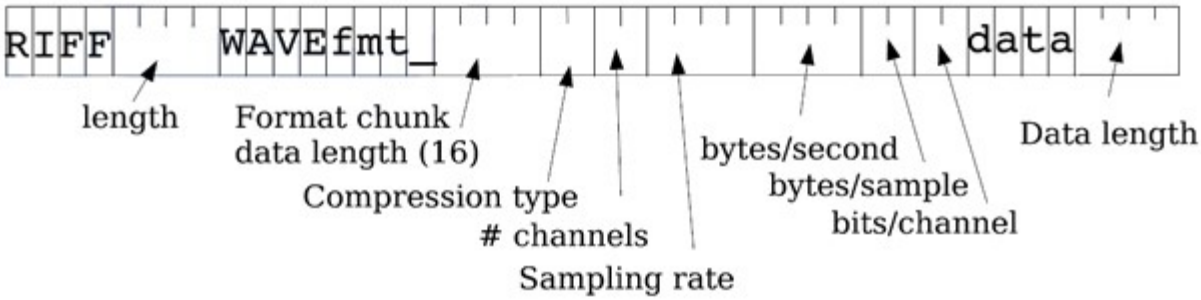
Prvi korak u obradi govora je pretvaranje analognih signala (prvenstveno tlaka zraka, a zatim analognog električnog signala u mikrofonu), u digitalni prikaz, odnosno digitalni signal. Ovaj proces analogno-digitalne pretvorbe ima dva koraka: uzorkovanje i kvantizaciju (Jurafsky & Martin, 2019).

Signal se uzorkuje mjerenjem njegove amplitude u određeno vrijeme, a brzina uzorkovanja je broj uzoraka uzetih u sekundi. Kako bi se valovi točno izmjerili potrebno je imati najmanje dva uzorka u svakom ciklusu: jedan koji mjeri pozitivni dio vala i jedan koji mjeri negativni dio vala. Više od dva uzorka po ciklusu povećavaju točnost mjerenja amplitude, ali manje od dva uzorka uzrokovat će potpuno propuštanje frekvencije vala. Maksimalni frekvencijski val koji se može mjeriti je onaj čija je frekvencija polovica stope uzorkovanja (budući da su za svaki

ciklus potrebna dva uzorka). Ova minimalna frekvencija s kojom je moguće uzorkovati analogni signal naziva se Nyquistova frekvencija. Većina informacija u ljudskom govoru je često frekvencije ispod 10000 Hz, stoga bi za razumijevanje i maksimalnu točnost bila potrebna brzina uzorkovanja od 20000 Hz. Telefonski govor filtrira komutacijska mreža (engl. *switching network*) te se samo frekvencije manje od 4000 Hz prenose telefonima. Stopa uzorkovanja od 8000 Hertza je dovoljna za telefonski govor. Brzina uzorkovanja od 16000 Hz (ponekad nazvana širokopojasnom) često se koristi za govor mikrofona. Čak i brzina uzorkovanja od 8000 Hz zahtijeva 8000 mjerenja amplitude za svaku sekundu govora, pa je važno učinkovito pohraniti mjerenje amplitude. Obično se pohranjuju kao cijeli brojevi (engl. *integers*), bilo 8-bitni (vrijednosti od -128 do 127) ili 16-bitni (vrijednosti od -32768 do 32767) (Jurafsky & Martin, 2019).

Ovaj proces opisivanja realnih brojeva cijelim brojevima zove se kvantizacija jer postoji minimalna granularnost (kvantna veličina) između pojedinih vrijednosti, i sve vrijednosti koje su bliže jedna drugoj od ove kvantne veličine predstavljene su identično. Jednom kad su podaci kvantizirani, pohranjuju se u različitim formatima. Parametri formatiranja kvantiziranih govornih signala su stopa uzorkovanja i veličina uzorka. Telefonski govor često je uzorkovan na 8000 Hz i pohranjen kao 8-bitni uzorak, dok se podaci mikrofona često uzorkuju na 16000 Hz i pohranjuju kao 16-bitni uzorci. Još jedan parametar pri formatiranju kvantiziranih govornih signala je broj kanala. Za stereo podatke ili za dvostrane razgovore, možemo pohraniti oba kanala u istoj datoteci ili ih možemo pohraniti u zasebne datoteke. Primjerice, uobičajeni format kompresije koji se koristi za telefonski govor je μ-zakon. Algoritmi za kompresiju zapisa poput μ-zakona uračunavaju osjetljivost ljudskog sluha na nižim frekvencijama, odnosno zapis niže frekvencije i vrijednosti zapisuje detaljnije od viših frekvencija i vrijednosti (Jurafsky & Martin, 2019).

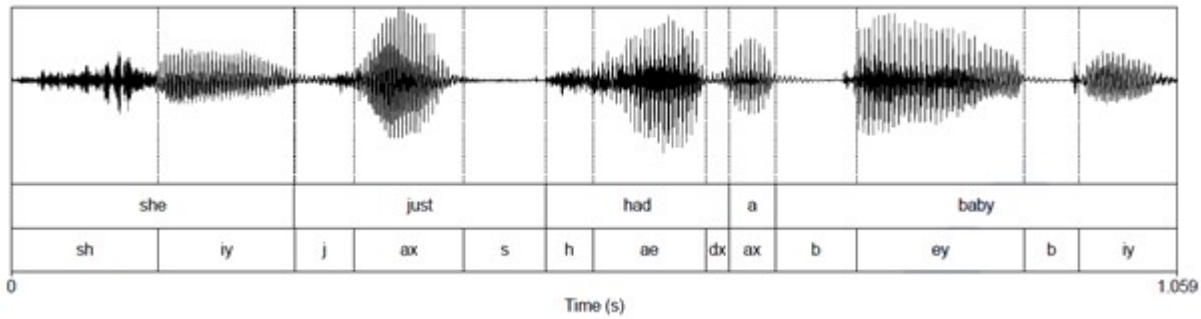
Postoji niz standardnih formata datoteka za pohranjivanje digitalizirane datoteke glasovnog signala, kao što su WAV, AIFF i AU, a svi imaju posebna zaglavlja, u upotrebi su također jednostavne 'sirove' (engl. *raw*) datoteke bez zaglavlja. Na primjer, .wav format je podskup Microsoftovog RIFF formata za multimedijske datoteke. RIFF je opći format koji može predstavljati niz ugniježđenih dijelova podataka i kontrolnih informacija. Slika 7 prikazuje jednostavnu .wav datoteku s jednim blokom podataka zajedno s blokom formata (Jurafsky & Martin, 2019).



Slika 7 - zaglavlje .wav formata

5.2.1. *Prepoznavanje fonema*

Vizualni pregled valnog oblika omogućuje puno saznanja o zapisanom govornom signalu. Na primjer, samoglasnike je prilično lako uočiti. Podsjetimo se da su samoglasnici zvučni; drugo svojstvo samoglasnika je da su obično dugi i relativno su glasni. Duljina se u vremenu očituje izravno na x-osi, dok se glasnoća prikazuje amplitudom na y-osi. Slika 8 prikazuje valni oblik kratke fraze na engleskom jeziku „*She just had a baby*." (Jurafsky & Martin, 2019).



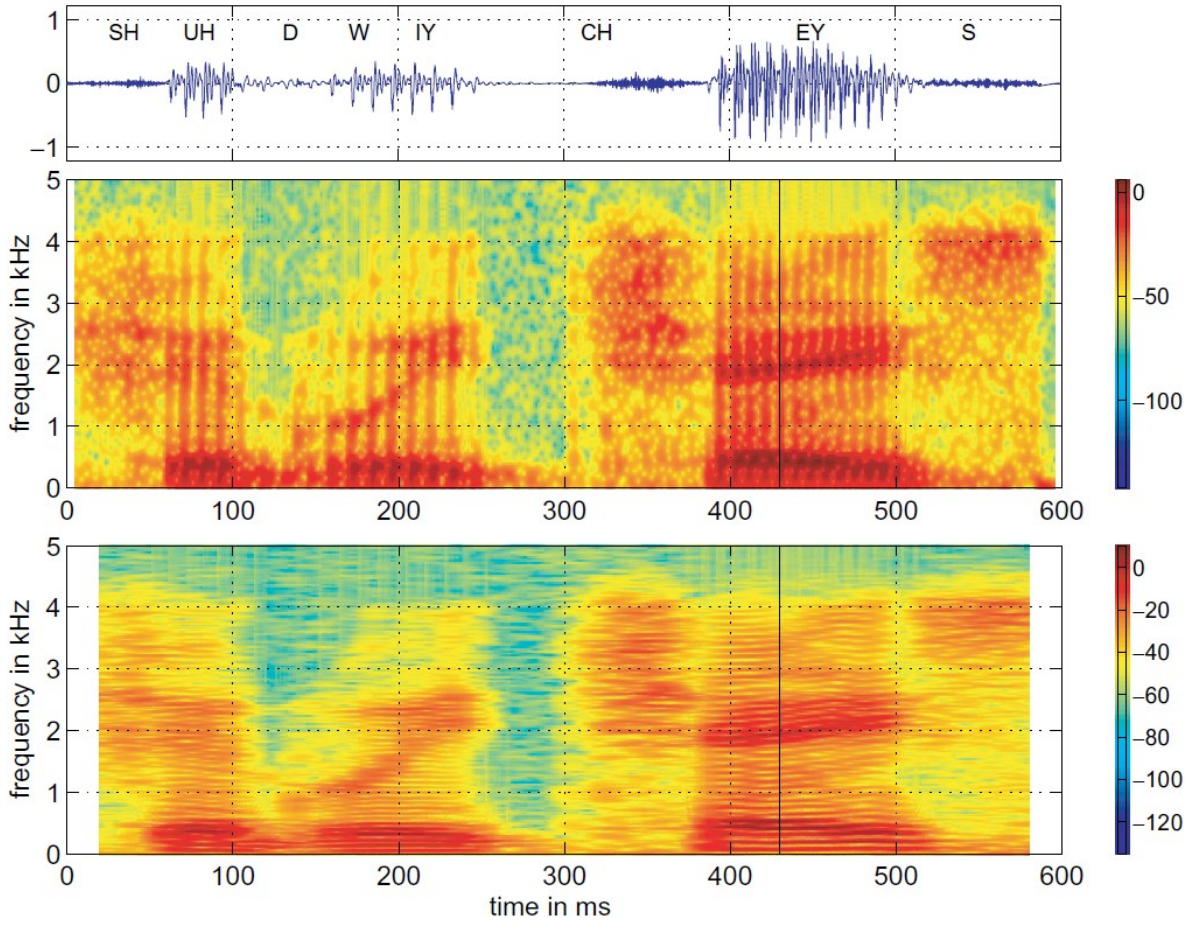
Slika 8 - valni prikaz rečenice "She just had a baby." (Jurafsky & Martin, 2019)

Valni oblik označen je oznakama riječi i fonema. Svi samoglasnici engleskog jezika u primjeru ([iy], [ax], [ae], [ax], [ey], [iy]) imaju pravilne vrhove amplitude koji ukazuju na zvučnosti. Za zaustavni suglasnik, koji se sastoji od zatvaranja nakon kojeg slijedi otpuštanje, možemo često vidjeti razdoblje tišine ili gotovo tišine nakon čega slijedi blagi nalet amplitude, kao što je vidljivo u oba [b] u riječi „*baby*". Također lako prepoznatljivi fonem je frikativ. Frikativi su vrlo oštri, poput [sh] te su izvedeni kad uzak kanal za protok zraka uzrokuje bučan i turbulentan zrak. Rezultirajući zvukovi imaju bučan i nepravilan oblik (Jurafsky & Martin, 2019).

5.3. Govorni spektrogrami

Od 1940-ih, spektrogram zvuka je osnovni alat za razumijevanje kako se proizvode zvukovi govora i kako se fonetičke informacije kodiraju u govoru. Do 1970-ih, spektrogrami su se izrađivali pomoću uređaja koji se sastojao od audio vrpce u petlji, varijabilnog analognog propusnog filtra i elektromehanički osjetljivog papira. Danas se spektrogrami poput

spektrograma prikazanog na slici 9 izrađuju tehnikama digitalne obrade signala i prikazuju se kao slike u bojama ili sivim tonovima na računalnim zaslonima (Rabiner & Schafer, 2007).



Slika 9 - spektrogram rečenice „Should we chase“ (Rabiner & Schafer, 2007)

Slika 9 prikazuje na vrhu valni oblik govornog signala rečenice na engleskom jeziku „*Should we chase*“, a ispod su prikazana dva spektrograma izračunata s različitim duljinama analitičkih prozora. Trake s desne strane kalibriraju boje spektrograma u decibelima (Rabiner & Schafer, 2007).

Pažljivo tumačenje valnih oblika govora i odgovarajućih spektrograma daje koristan uvid u prirodu govornog signala. Duljina prozora ima značajan utjecaj na spektrogram. Gornji spektrogram na slici 9 izračunat je s duljinom prozora od 101 uzorka, što odgovara trajanju od 10 milisekundi. Kao rezultat toga, u glasovnim intervalima spektrogram prikazuje uspravno orijentirane pruge koje odgovaraju činjenici da glasovni signal u vremenskom prozoru ponekad uključuje uglavnom uzorke velike amplitude, a ponekad uglavnom uzorke male amplitude. Kao rezultat kratke duljine prozora analize, svaki pojedinačni period visine tona je jasno razlučen u vremenskoj dimenziji, ali je razlučivost (engl. *resolution*) u dimenziji frekvencije loša. Iz tog

razloga, ako je analitički prozor kratak, spektrogram se naziva širokopojasni spektrogram (Rabiner & Schafer, 2007).

Nasuprot tome, kada je duljina prozora velika, spektrogram je uskopojasni spektrogram, koji se odlikuje dobrom frekvencijskom rezolucijom i lošom razlučivošću u vremenskoj dimenziji, što se može vidjeti na donjem spektrogramu. Donji spektrogram na slici 9 izračunat je s duljinom prozora od 401 uzorka, što odgovara trajanju od 40 milisekundi. S obzirom na to, spektrogram više ne prikazuje vertikalno orijentirane pruge budući da je nekoliko vremenskih razdoblja uključeno u prozor bez obzira gdje se nalazi na valnom obliku u blizini vremena analize. Iz tog razloga, spektrogram nije tako osjetljiv na brze vremenske varijacije, ali je rezolucija u frekvencijskoj dimenziji mnogo bolja. Stoga su pruge u slučaju uskopojasnog spektrograma horizontalno orijentirane jer su temeljna frekvencija i njeni harmonici razlučeni (Rabiner & Schafer, 2007).

6. Analiza govornog signala

Postoje tri glavne faze u prepoznavanju govora, a to su analiza signala, ekstrakcija (izdvajanje) značajki i modeliranje. Izdvajanje značajki igra važnu ulogu u sustavu prepoznavanja govora i dobra tehnika izdvajanja značajki govora omogućit će sustavima da točno identificiraju izgovorene riječi (Zulkifly & Yahya, 2017). U prošlosti, sustavi prepoznavanja govornog signala obično su koristili mel-frekvencijski kepstralni koeficijent (engl. *mel-frequency cepstral coefficient*, MFCC) ili relativno spektralno transformiranje - perceptualno linearno predviđanje (engl. *relative spectral transform-perceptual linear prediction*, RASTA-PLP) kao vektore značajki i Gaussov model mješavine kao akustički model (Yu & Deng, 2015). Mel-frekvencijski kepstralni koeficijent se računa prolazeći kroz pet procesa kojima se iz govornog signala izvode informacije koje se čuju ljudskim uhom, odnosno izdvajaju vokalne značajke govornog signala te se koristi za prepoznavanje govornika u sustavu prepoznavanja govora (Abdul & Al-Talabani, 2022). Relativno spektralno transformiranje - perceptualno linearno predviđanje koristi se kao način iskrivljenja spektra kako bi se minimizirale razlike između govornika uz očuvanje važnih govornih informacija (Zulkifly & Yahya, 2017).

6.1. Model Gaussovih mješavina

Model Gaussovih mješavina (engl. *Gauss mixture model*, GMM) je statistički model koji predstavlja distribuciju vjerojatnosti na određenom skupu podataka te je vrlo važan u akustičkom modeliranju sustava prepoznavanja govora. Model Gaussovih mješavina radi tako da procjenjuje parametre tih Gaussovih distribucija, uključujući njihove srednje vrijednosti i

varijance. Također dodjeljuje težinu svakoj Gaussovoj distribuciji, koja predstavlja važnost ili vjerojatnost prisutnosti svake grupe u skupu podataka. Za pronalazak najbolje prilagođenog modela Gaussovih mješavina za određeni skup podataka, model iterativno prilagođava parametre i težine koristeći algoritam nazvan algoritam maksimizacija očekivanja (engl. *Expectation-Maximization*, EM). Tijekom koraka očekivanja, procjenjuje se vjerojatnost da svaka točka podataka pripada svakoj Gaussovoj distribuciji. U koraku maksimizacije, ažuriraju se parametri na temelju tih vjerojatnosti. Kada je model Gaussovih mješavina istreniran, može se koristiti za prognoziranje. Za novu točku podataka, model izračunava vjerojatnost da pripada svakoj Gaussovoj distribuciji i dodjeljuje je grupi s najvećom vjerojatnošću. Model Gaussovih mješavina pomaže u identificiranju različitih grupa unutar skupa podataka pretpostavljajući da svaka grupa slijedi krivulju u obliku zvona. Uči karakteristike tih grupa i koristi to znanje za klasifikaciju novih točaka podataka na temelju najvjerojatnije grupe kojoj pripadaju (Yu & Deng, 2015).

6.2. Skriveni Markovljev model

Zbog fizičkih ograničenja, ljudski glas i govor ne mogu proizvesti drastične promjene te postoje kratki intervali gdje svi organi potrebni za artikulaciju ostaju relativno stacionarni (Chou & Juang, 2003). Skriveni Markovljev model (engl. *hidden Markov model*, HMM) je statistički model koji se koristi za modeliranje niza događaja, a ključna ideja skrivenog Markovljevog modela je da je svaki događaj u nizu povezan sa „skrivenim“ stanjem koje ga generira, a mi promatramo samo događaj, a ne stanje. Stanje je samo po sebi slučajna varijabla koja obično ima diskretne vrijednosti. Skriveni Markovljev model može se zamisliti kao stroj koji se nalazi u jednom od nekoliko mogućih stanja, pri čemu svako stanje ima određenu vjerojatnost prijelaza u drugo stanje. Na temelju trenutnog stanja, skriveni Markovljev model generira događaj (npr. riječ ili simbol) s određenom vjerojatnošću. Zatim se prelazi u novo stanje, i taj proces se ponavlja za svaki događaj u nizu (Yu & Deng, 2015). Ključni izazov skrivenog Markovljevog modela je određivanje vjerojatnosti prijelaza između stanja i vjerojatnosti generiranja događaja za svako stanje. Ovo se obično radi treniranjem modela na temelju postojećih podataka, koristeći algoritme poput Baum-Welch algoritma ili Viterbi algoritma (Chou & Juang, 2003).

6.3. Prepoznavanje govora korištenjem umjetnih neuronskih mreža

U 80-im godinama prošlog stoljeća pojavila se nova paradigma umjetnih neuronskih mreža. Neuronske mreže razlikuju se od prethodno prevladavajućih modela prepoznavanja govora po tome što imaju puno veću mogućnost razlikovanja i razlučivanja (Chou & Juang, 2003).

Posljednjih godina diskriminativni hijerarhijski modeli kao što su duboke neuronske mreže (engl. *deep neural network*, DNN) postali su izvedivi zahvaljujući tehnološkom napretku i značajno su smanjili stope pogrešaka, zahvaljujući stalnim poboljšanjima u računalnoj snazi, dostupnosti velikog skupa podataka za treniranje modela i boljem razumijevanju ovih modela. Na primjer, kontekstno ovisna hibridna duboka neuronska mreža sa skrivenim Markovljevim modelom (CD)-DNN-HMM postigla je jednu trećinu smanjenja stope pogrešaka na zadatku govorne transkripcije u odnosu na konvencionalne sustave modela Gaussovih mješavina i skrivenih Markovljevih modela (Yu & Deng, 2015).

Sustav neuronskih mreža obično se sastoji od mrežne strukture i mnogo jednostavnih operacijskih jedinica. Tako je sustav neuronskih mreža distribuiran sustav paralelnih računalnih mehanizama što je povoljno za visoku brzinu i robusnost upotrebe sustava. Sustavi neuronskih mreža zahvaljujući svojoj robusnosti imaju visok stupanj otpornosti na grešku i visoku stabilnost i pouzdanost kod klasifikacijskih zadataka. Uspjeh sustava neuronskih mreža ovisi izravno o odabiru procedura za treniranje sustava i skupa podataka kojima se sustav trenira (Chou & Juang, 2003).

Govorni signali su uglavnom dinamični dok je osnovna struktura sustava neuronske mreže stvorena za rukovanje statičnim uzorcima. Zbog tog problema provodila su se istraživanja kako bi se riješila neusklađenost između dinamike govornih signala i uzoraka potrebnih za korištenje neuronskih mreža te su se koristili kratki segmenti govornih signala. Kasnije je razvijen već spomenut hibridni model sustava prepoznavanja govora koji koristi neuronske mreže uz pomoć skrivenog Markovljevog modela (Chou & Juang, 2003). U ovom hibridnom modelu je dinamika govornog signala modelirana pomoću skrivenog Markovljevog modela, a vjerojatnosti opažanja procijenjene su pomoću neuronskih mreža. Svaki izlazni neuron umjetne neuronske mreže osposobljen je za procjenu vjerojatnosti stanja skrivenog Markovljevog modela s obzirom na akustička opažanja. Hibridni modeli neuronskih mreža i skrivenog Markovljevog modela imaju dodatnu prednost što je dekodiranje govornog signala općenito učinkovito (Yu & Deng, 2015).

7. Prepoznavanje govora za hrvatski jezik

U posljednjem desetljeću je obrada prirodnog jezika postigla iznimne rezultate i brz razvoj. Iz perspektive korisnika, to se odražava u uslugama i aplikacijama koje pružaju pouzdano prepoznavanje govora i sintezu prirodnog govora, kao i napredno razumijevanje prirodnog jezika i generiranje prirodnog jezika. Takav se napredak može zahvaliti i činjenici da su uređaji

krajnjih korisnika značajno evoluirali: moćni pametni telefoni i razni drugi pametni uređaji sada su široko dostupni. Poboljšana je prateća infrastruktura što je omogućilo već spomenuti razvoj u obradi prirodnog jezika, no sva navedena poboljšanja uglavnom se pripisuju engleskom jeziku. Iz perspektive manjinskog jezika kao što je hrvatski, još uvijek postoje mnogi izazovi za postizanje napretka usporedivog s uslugama dostupnim za engleski jezik (Šoić & Vuković, 2022). Što se govornih tehnologija tiče, hrvatski jezik je još uvijek nedovoljno razvijen pogotovo u pogledu prepoznavanja govora, iako je od posljednje temeljite analize prošlo gotovo deset godina (Ipšić & Martinčić-Ipšić, 2010). Očigledni razlog ovakvog stanja je činjenica da nije relevantan na globalnoj razini. Manjinski jezici rijetko su dobro razvijeni u području govornih tehnologija jer nemaju značajne ekonomske koristi. Drugi razlog povezan je s jezičnom složenošću. Hrvatski je jezik morfološki vrlo bogat, što povećava složenost i povećava broj izazova u mnogim područjima obrade prirodnog jezika. Metode koje se koriste u slučaju engleskog jezika ne mogu se izravno primijeniti jer je hrvatski jezik različit u ortografskom, fonetičkom i morfološkom smislu. Osim toga, još uvijek ne postoje prikladni otvoreni skupovi podataka za potrebe istraživanja (Šoić & Vuković, 2022).

Jedan od postojećih sustava za prepoznavanje govora hrvatskog jezika temelji se na kontinuiranim skrivenim Markovljevim modelima kontekstualno neovisnih (monofoni) i kontekstualno ovisnih (trofoni) akustičkih modela. Treniranje sustava za prepoznavanje govora provedeno je pomoću alata HTK (engl. *Hidden Markov Model Toolkit*) (Ipšić & Martinčić-Ipšić, 2010). Hidden Markov Model Toolkit (HTK) alat je za izradu i manipuliranje skrivenim Markovljevim modelima. HTK se primarno koristi za istraživanje prepoznavanja govora iako se koristi za brojne druge primjene uključujući istraživanje sinteze govora, prepoznavanja znakova i sekvenciranja DNK (HTK, 2016). Ipšić i Martinčić-Ipšić su u svom radu pokazali da je pristup za prepoznavanje govora korištenjem akustičkog modeliranja ovisnog o kontekstu prikladan za brzi razvoj govornih aplikacija ograničene domene za jezike s malo resursa poput hrvatskog. Predložena su se hrvatska pravopisno-fonetska pravila za izgradnju fonetskoga rječnika. Razvijeni hrvatski govorni korpus s više govornika uspješno je korišten za razvoj govornih aplikacija. Predložena hrvatska fonetska pravila obuhvatila su odgovarajuće fonetsko, jezično i artikulacijsko znanje hrvatskog jezika za vezivanje stanja u akustičkim modelima sustava za prepoznavanje govora. Glavna prednost korištenog pristupa leži u činjenici da se govorne aplikacije mogu učinkovito i brzo prenijeti na druge interesne domene pod uvjetom da postoji odgovarajući govorni i jezični korpus. Budući da je bio planiran telefonski pristup govornom dijaloškom sustavu, bilo je potrebno razmotriti daljnja poboljšanja u prepoznavanju

govora. Osim toga, u tijeku je bio rad na uključivanju više govora, posebno spontanog govora različitih govornika u korpus. Daljnje istraživačke aktivnosti također su bile planirane prema razvoju modula za razumijevanje govora u dijaloškom sustavu i modula za sintezu govora (Ipšić & Martinčić-Ipšić, 2010).

7.1. Postojeći alati za prepoznavanje govora za hrvatski jezik

Postoje mnogi, često komercijalni alati za prepoznavanje govora na hrvatskom jeziku i pretvorbu govora u tekst. U ovom poglavlju bit će navedeni neki od popularnijih i više korištenih alata, kao i način na koji odrađuju pretvorbu govora u tekst.

CMUSphinx je otvorena biblioteka za prepoznavanje govora koja podržava nekoliko jezika, uključujući hrvatski. Pruža Python API (engl. *application programming interface*) za jednostavnu integraciju u Python projekte. CMUSphinx koristi skup modela koji su trenirani i temeljeni na akustičkoj i jezičnoj statistici kako bi prepoznao govor i pretvorio ga u tekst (CMUSphinx, 2023).

Google Cloud Speech-to-Text je usluga koja pruža napredno prepoznavanje govora u stvarnom vremenu. Iako izvorno nije usmjeren samo na hrvatski jezik, Google Cloud Speech-to-Text podržava veliki broj jezika, uključujući i hrvatski. Pruža mogućnosti prepoznavanja i visoku razinu preciznosti (Google, 2023).

Mozilla DeepSpeech je otvoreni projekt razvijen od strane Mozille koji koristi neuronske mreže za prepoznavanje govora. DeepSpeech ima podršku za hrvatski jezik te može pretvoriti govor u tekst. Ovaj alat zahtijeva treniranje modela na velikim skupovima podataka kako bi postigao visoku razinu preciznosti (Mozilla Corporation, 2023).

Kaldi je popularan otvoreni sustav za prepoznavanje govora koji podržava hrvatski jezik. Kaldi je fleksibilan alat koji se često koristi u akademskim i istraživačkim krugovima. Omogućava prilagodbu i podešavanje modela prepoznavanja govora za poboljšanu preciznost (Kaldi, 2023).

Wit.ai je platforma za razvoj chatbotova koja također pruža podršku za prepoznavanje govora. Omogućuje integraciju s Pythonom putem svog Python SDK-a (engl. *software development kit*). Wit.ai platforma koristi duboko učenje i strojno učenje kako bi prepoznala govorne ulazne signale i konvertirala ih u tekst (wit.ai, 2023).

Microsoft Azure Speech Services je još jedna popularna usluga za prepoznavanje govora. Podržava razne jezike, uključujući hrvatski, i omogućuje pretvaranje govora u tekst putem API-ja ili lokalno korištenjem SDK-ova (Microsoft Azure, 2023).

Uz navedene alate postoje i mnoge druge web i mobilne aplikacije koje mogu poslužiti kao dobar početak za prepoznavanje govora i pretvorbu govora u tekst na hrvatskom jeziku. Važno je pri upotrebi alata uzeti u obzir njihove značajke, performanse i prilagoditi ih prema specifičnim potrebama projekta.

8. Analiza algoritama za hrvatski jezik u Pythonu

Python je interpretirani programski jezik visoke apstrakcijske razine poznat po svojoj jednostavnosti i čitljivosti. Kreirao ga je Guido van Rossum i prvi put je objavljen 1991. Python naglašava čitljivost koda i ima za cilj pružiti jasnu i konciznu sintaksu, olakšavajući programerima da izraze svoje ideje i učinkovito razvijaju softver. Python je jezik koji ima sintaksu laku za čitanje, tj. koristi čistu i jednostavnu sintaksu što olakšava razumijevanje i pisanje koda. Python programi se izvode red po red pomoću Python interpretera, što omogućuje brzi razvoj i testiranje. Također podržava interaktivni način rada, gdje se mogu izravno unijeti naredbe i dobiti trenutni rezultati. Python je dostupan na više platformi kao što su Windows, macOS i razne distribucije Linuxa, što programerima omogućuje pisanje koda na jednoj platformi i pokretanje na drugoj bez većih izmjena. Python dolazi s opsežnom standardnom bibliotekom koja pruža širok raspon modula i funkcija za razne zadatke. Standardna biblioteka smanjuje potrebu programera da pišu kod od nule, jer su mnoge uobičajene operacije već implementirane. Python ima golemu zajednicu programera koji pridonose njegovom rastu i razvoju. To je rezultiralo bogatim ekosustavom biblioteka i okvira trećih strana koji proširuju mogućnosti Pythona, kao što su Django, NumPy, Pandas, TensorFlow i mnogi drugi. Popularnost Pythona znatno je porasla tijekom godina i postao je jedan od najčešće korištenih programskih jezika u raznim domenama. Njegova jednostavnost, čitljivost i opsežan ekosustav čine ga odličnim izborom i za početnike i za iskusne programere (Python Software Foundation, 2023).

Python SpeechRecognition popularna je biblioteka koja programerima omogućuje ugradnju funkcije prepoznavanja govora u svoje Python aplikacije. Djeluje kao most između programskog jezika Python i raznih mehanizama za prepoznavanje govora i API-ja, omogućujući korisnicima transkripciju govornog jezika u pisani tekst. Jedna od ključnih prednosti Python biblioteke SpeechRecognition je njezina jednostavnost i laka upotreba. Omogućuje jednostavno i intuitivno sučelje za rad s funkcijama prepoznavanja govora. Sa samo nekoliko redaka koda, programeri mogu „uhvatiti“ audio ulaz, poslati ga mehanizmu za prepoznavanje govora i primiti transkribirani tekst kao izlaz. Biblioteka podržava više

mehanizama za prepoznavanje govora, uključujući Google Speech Recognition, CMU Sphinx i Microsoft Azure Speech. Ovi sustavi (engl. *engines*) koriste napredne algoritme i tehnike strojnog učenja za analizu i interpretaciju govornih podataka, postižući visoku točnost i performanse. Pythonova biblioteka SpeechRecognition nudi niz značajki i mogućnosti. Podržava online i offline prepoznavanje govora, omogućujući korisnicima transkripciju audio ulaza u stvarnom vremenu ili obradu unaprijed snimljenih audio datoteka. Osim toga, pruža opcije za podešavanje parametara prepoznavanja, kao što su odabir jezika, konfiguracija izvora zvuka i rukovanje šumom, kako bi se poboljšala točnost u određenim scenarijima. Kompatibilnost biblioteke s različitim platformama i operativnim sustavima, uključujući Windows, macOS i Linux, čini je svestranom i dostupnom širokom rasponu programera. Također se dobro integrira s drugim Python bibliotekama i okvirima, omogućujući besprijekornu integraciju u postojeće projekte. Pythonova biblioteka SpeechRecognition naširoko se koristi u različitim aplikacijama, kao što su glasovni pomoćnici, usluge transkripcije i glasovno kontrolirani sustavi. Njena jednostavnost, opsežna dokumentacija i aktivna podrška zajednice čine ga popularnim izborom za programere koji žele uključiti mogućnosti prepoznavanja govora u svoje Python projekte (Python Software Foundation, 2023).

U nastavku rada bit će izloženi rezultati testiranja Python algoritama za prepoznavanje govora, fokusirajući se posebno na hrvatski jezik. Cilj je procijeniti točnost i učinkovitost ovih algoritama prije njihove potrebe za dodatnim kalibriranjem korištenjem muškog i ženskog glasa uz korištenje identičnih skupova rečenica. Odabrane rečenice ulomci su iz prvog poglavlja dječje knjige „Vlak u snijegu“ Mate Lovraka.

Za provedbu testiranja bit će korišteno nekoliko biblioteka za prepoznavanje govora, alati dostupni u Pythonu koji podržavaju hrvatski jezik i moduli potrebni za korištenje biblioteka i alata. Za svaki od korištenih alata bit će predstavljeni dodatni koraci koje je potrebno poduzeti pri testiranju. Koristeći ove resurse bit će uspoređena izvedba i pouzdanost različitih algoritama u točnom prepisivanju izgovorenih riječi na hrvatskom jeziku. Pri testiranju neće biti uključeno prepoznavanje interpunkcije.

Proces testiranja uključivat će predstavljanje istih unaprijed određenih rečenica izrečenih muškim i ženskim glasom, osiguravajući dosljedne varijable kao što su ton, brzina i izgovor. Bit će upotrijebljeni muški i ženski glas kako bi se testirala osjetljivost postojećih tehnologija na razlike u visini tona između muškog i ženskog glasa. Održavanjem kontroliranog okruženja, sve razlike u točnosti prepoznavanja između glasova moći će se pripisati isključivo temeljnim algoritmima. Datoteke zvučnih zapisa su u .wav formatu, datoteka koja sadrži zvučni zapis

muškog glasa je veličine 4.627 KB, dok je datoteka koja sadrži zvučni zapis ženskog glasa veličine 4.301 KB.

Budući da Python SpeechRecognition biblioteka sadrži više mehanizama za prepoznavanje govora, u nastavku će biti testirani Google Speech Recognition, Wit.ai i Google Cloud Services Speech-to-Text.

8.1. Google Speech Recognition

Google Speech Recognition naširoko je korištena usluga prepoznavanja govora koju je razvio Google. Programerima omogućuje pretvaranje govornog jezika u pisani tekst, omogućujući aplikacijama obradu i razumijevanje ljudskog govora. Google Speech Recognition koristi algoritme strojnog učenja i veliku količinu podataka za treniranje kako bi pružio točnu i pouzdanu transkripciju govora u tekst. Jedna od ključnih prednosti Google prepoznavanja govora je njegova robusnost. Treniran je na korpusu višejezičnih podataka, što ga čini sposobnim za prepoznavanje govora na brojnim jezicima i naglascima. Bilo da se radi o kratkoj frazi ili dugoj zvučnoj snimci, Google Speech Recognition djeluje u različitim slučajevima upotrebe, uključujući usluge prijepisa, glasovne pomoćnike i još mnogo toga. Google Speech Recognition također nudi mogućnosti prepoznavanja govora u stvarnom vremenu, omogućujući trenutnu transkripciju zvuka dok se izgovara. Ova je značajka posebno korisna za aplikacije koje zahtijevaju trenutnu povratnu informaciju ili transkripciju uživo, kao što su sustavi s glasovnom kontrolom ili usluge transkripcije u stvarnom vremenu. Google Speech Recognition pruža API jednostavan za korištenje koji programeri mogu integrirati u svoje aplikacije. API nudi i sinkrone i asinkrone metode prepoznavanja govora, omogućujući programerima da odaberu najprikladniju opciju na temelju svojih specifičnih zahtjeva (Python Software Foundation, 2023).

Za potrebe ovog rada, Google Speech Recognition uvezen je u Python kod putem biblioteke SpeechRecognition i uz pomoć modula uvedena je audio datoteka. Metoda recognize_google kojom se provodi prepoznavanje govora i pretvorba govora u tekst u predstavljenom odlomku koda 1 ima dva argumenta. Prvi od tih argumenata je audio datoteka te je on obavezan, a drugi, koji je u ovom slučaju jezik, nije obavezan, ali je postavljen za kvalitetniju primjenu predefiniranih modela.

Odlomak koda 1 - Google Speech Recognition

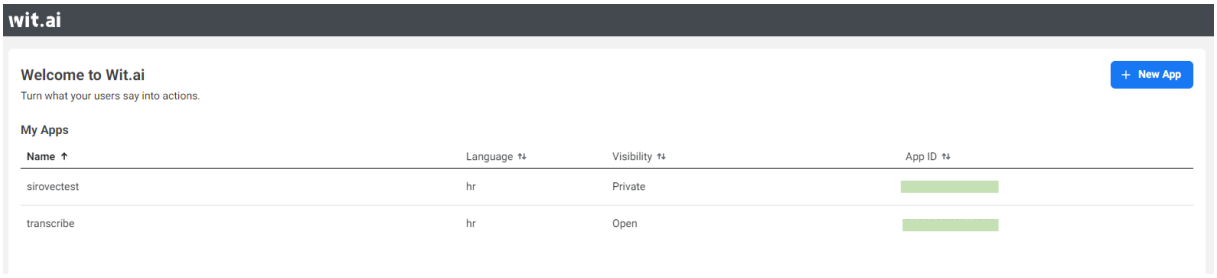
```
1 import speech_recognition as sr
2 r=sr.Recognizer()
3 audio_file = "lovrak.wav"
4 with sr.AudioFile(audio_file) as source:
5     audio = r.record(source)
6 text = r.recognize_google(audio, language='hr-HR')
7 print("Transkripcija: " + text)
```

8.2. Wit.ai

Facebook je izradio platformu za razumijevanje prirodnog jezika (engl. *natural language understanding*, NLU) pod nazivom Wit.ai koja pomaže programerima u stvaranju alata koji učinkovito razumiju i prevode ljudski jezik. Moguće je koristiti Wit.ai za izgradnju sučelja za razgovor, chatbota, glasovnih asistenata i još mnogo toga korištenjem raznih alata za obradu jezika. Primarni cilj Wit.ai-ja je „izvući“ značenje iz korisničkih unosa, uključujući tekst i zvuk. Programeri mogu odrediti namjere, koje ukazuju na cilj ili svrhu korisnika, i entitete, koji predstavljaju značajne dijelove informacija unutar korisničkog unosa, koristeći njegovo jednostavno sučelje prilagođeno korisniku. Wit.ai trenira modele pomoću tehnika strojnog učenja kako bi mogli točno prepoznati i izdvojiti namjere i entitete iz korisničkih upita. Wit.ai-jeva sposobnost da se nosi s promjenama prirodnog jezika i razumije kontekst, jedna je od njegovih glavnih prednosti. Omogućuje razvijanje novih entiteta koji su specifični za aplikacijsku domenu, dok također pruža ugrađene vrste entiteta kao što su brojevi, datumi, trajanja i više. Dodatno, Wit.ai nudi pomoć pri rukovanju postavkama razgovora, omogućujući programu da sačuva stanje i isporuči uvjerljivije i interaktivnije odgovore. Moguće je kreirati i trenirati vlastite jezične modele pomoću korisničkog web sučelja koje nudi Wit.ai. Nudi niz alata za pregled, testiranje i poboljšanje vlastitih modela. Nadalje, Wit.ai nudi RESTful API koji olakšava ugradnju Wit.ai značajki u vlastite aplikacije. Wit.ai je postao popularan za stvaranje konverzacijskih aplikacija i chatbotova zbog svoje prilagodljivosti i jednostavnosti korištenja. Prilagodljiv je i dostupan širokom rasponu programera zahvaljujući podršci za nekoliko programskih jezika, uključujući Python (wit.ai, 2023).

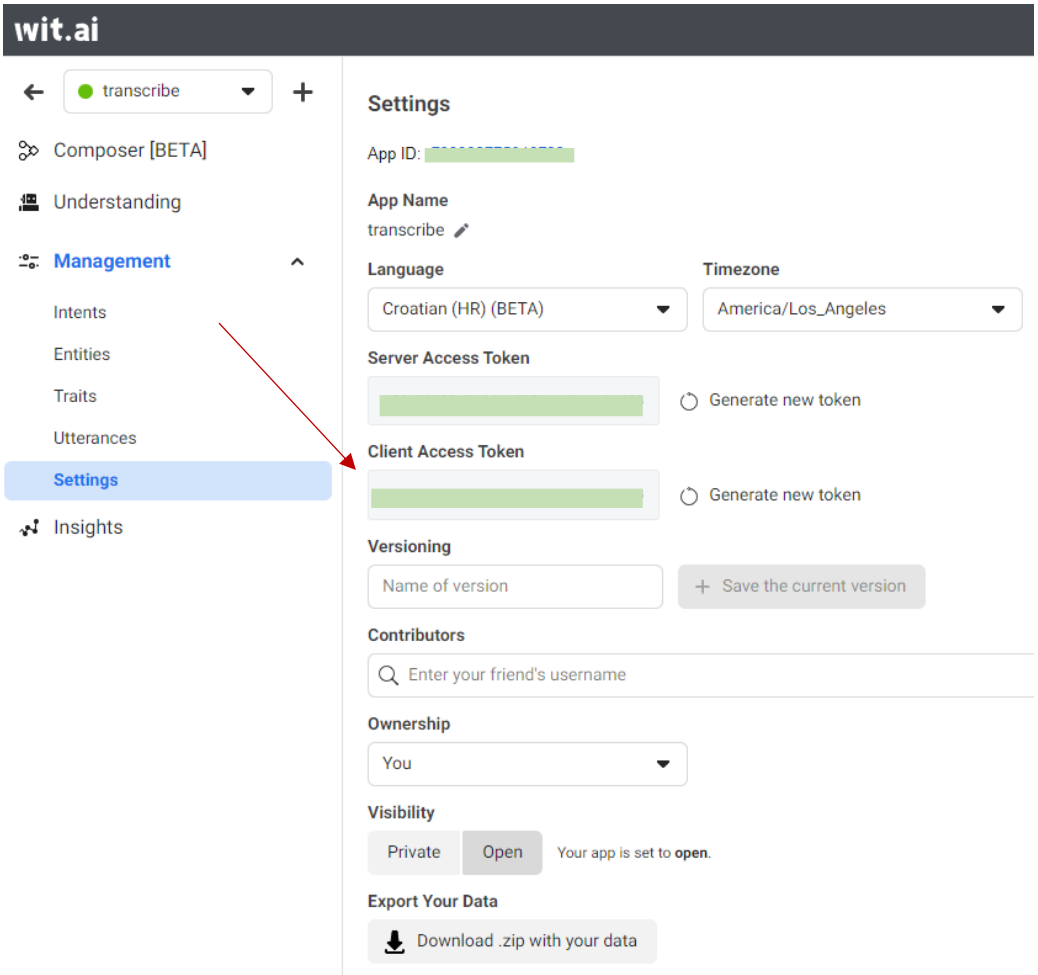
Za korištenje Wit.ai-ja bilo je potrebno ulogirati se u platformu putem Meta pristupnih podataka jer je za njegovo korištenje u Pythonu potreban klijentski pristupni token. Na slici 10 vidljiva

je naslovna stranica platforme Wit.ai. Za stvaranje klijentskog pristupnog tokena potrebno je napraviti novu aplikaciju pritiskom na gumb „+ New App“. Pri stvaranju aplikacije na platformi potrebno je odabrati željeni jezik, što je kasnije moguće vidjeti i promijeniti u postavkama aplikacije.



Slika 10 - naslovna stranica platforme Wit.ai

Nakon stvaranja nove aplikacije ona se pojavljuje na popisu aplikacija. Za pristup klijentskom tokenu potrebno je odabrati stvorenu aplikaciju i u izborniku s desne strane odabrati *Management* → *Settings* (Slika 11). Token je potom potrebno iskopirati i zalijepiti na odgovarajuće mjesto u Python kodu.



Slika 11 - postavke odabrane aplikacije u platformi Wit.ai

U odlomku koda 2 vidljivo je da je i za korištenje Wit.ai-a uvezena biblioteka SpeechRecognition kako bi se mogla uvesti audio datoteka. Prepoznavanje govora i pretvorba govora u tekst obavlja se metodom recongize_wit u kojoj su dva obavezna argumenta. Prvi je audio datoteka, a drugi je klijentski pristupni token koji je preuzet s web platforme Wit.ai.

Odlomak koda 2 - Wit.ai

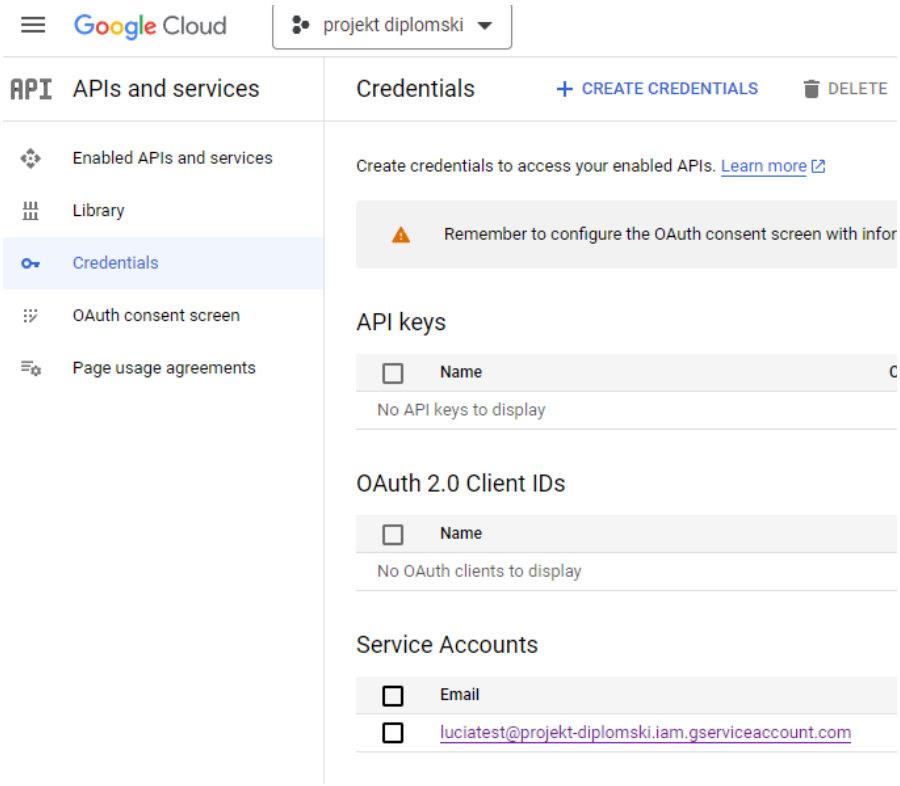
```
1 import speech_recognition as sr
2 access_token = '<token>'
3 r = sr.Recognizer()
4 with sr.AudioFile('lovrak.wav') as source:
5     audio = r.record(source)
6 text = r.recognize_wit(audio, key=access_token)
7 print('Transkripcija: ' + text)
```

8.3. Google Speech Cloud Services Speech-to-Text

Google Cloud Speech Recognition svestrana je usluga pretvorbe govora u tekst koju pruža Google Cloud Platform. Usluga se nakon probnog perioda naplaćuje te ju je moguće koristiti i bez mrežne povezanosti, za razliku od prije navedenog Google Speech Recognitiona te je njegovo svojevrsno proširenje. Koristi napredne tehnologije strojnog učenja za pretvaranje govornog jezika u pisani tekst, omogućujući programerima da u svoje aplikacije ugrade točne i učinkovite mogućnosti prepoznavanja govora. Uz Google Cloud Speech Recognition moguća je transkripcija zvuka iz različitih izvora, uključujući ulaz mikrofona, audio datoteke, ili strujanje podataka (engl. *streaming*). Google Cloud Speech-to-Text API podržava širok raspon audio formata i pruža mogućnosti za obradu u stvarnom vremenu i skupnu obradu za različite slučajeve upotrebe. Jedna od ključnih prednosti Google Cloud Speech Recognition Speech-to-Text API-ja je njegova točnost. Koristi Googleove opsežne jezične modele i goleme količine podataka za treniranje radi postizanja preciznih transkripcija, čak i u zahtjevnim audio uvjetima. Usluga može rukovati različitim jezicima, dijalektima i naglascima, što je čini prikladnom za globalne aplikacije. Štoviše, Google Cloud Speech Recognition nudi brojne značajke i mogućnosti prilagodbe. Podržava dijarizaciju govornika, što omogućuje prepoznavanje i razlikovanje govornika u razgovoru. Osim toga, pruža vremenske oznake na razini riječi, omogućujući precizno poravnavanje transkribiranog teksta s odgovarajućim audio segmentima. Google Cloud Speech-to-Text API također omogućuje određivanje specijaliziranih modela za

poboljšanu točnost prepoznavanja u određenim domenama, kao što su telefonija ili naredbe. Usluga je lako dostupna putem Google Cloud Speech-to-Text API-ja, koji programerima pruža jednostavan način integracije. Uz odgovarajuću provjeru autentičnosti i postavljanje, moguće je slati API zahtjeve za transkripciju zvuka i primati rezultate u strukturiranom i praktičnom formatu. Google Cloud Speech Recognition pronalazi aplikacije u raznim industrijama i slučajevima upotrebe. Može poboljšati pristupačnost pružanjem titlova za video sadržaj, poboljšati korisničku podršku automatskim prijepisom razgovora u pozivnom centru, olakšati analizu podataka pretvaranjem audio podataka u tekst koji se može pretraživati, i još mnogo toga. Iskorištavanjem snage Google Cloud Speech Recognition, programeri mogu povećati potencijal postojećih govornih podataka i izgraditi inovativne aplikacije koje nude besprijekorne i točne mogućnosti pretvaranja govora u tekst (Google, 2023).

Google Cloud Speech Recognition je od korištenih algoritama najkompleksniji za primjenu. Slično kao i kod korištenja alata Wit.ai, bilo je potrebno napraviti ključ za korištenje. U slučaju Google Cloud platforme, potrebno je ulogirati se u konzolu Google Clouda uz uvjet postojanja Google računa i stvoriti novi projekt. U projektu je potrebno odabrati API za korištenje, u ovom slučaju to je bio Speech. Nakon odabira API-ja, potrebno je stvoriti ključ za korištenje (Slika 12). U izborniku s desne strane potrebno je odabrati *Credentials* te stvoriti novi *Service Account*. Stvaranjem *Service Accounta* stvara se i JSON datoteka koja služi kao ključ za korištenje Google Clouda preko drugih platformi, u ovom slučaju u Pythonu.



Slika 12 - Google Cloud ključevi

Nakon izrade ključa potrebno je upisati njegovu adresu na računalu u Python kod. Za korištenje Google Cloud Speech-to-Text API-ja potrebno je uvesti više biblioteka u Python. To je naravno biblioteka SpeechRecognition, uz nju biblioteka Google Cloud i modul OS za upisivanje adrese pristupnog ključa Google Cloud API. Za korištenje Google Cloud Speech-to-Text API-ja potrebno je definirati funkciju kojom će biti odrađeno prepoznavanje govora i njegova pretvorba u tekst. Između ostalog, potrebno je odrediti konfiguraciju audio datoteke, tj. potrebno je odrediti jezik prepoznavanja i stopu uzorkovanja, što je vidljivo u odlomku koda 3. Stopa uzorkovanja ostavljena je na zadanoj vrijednosti, a jezik je postavljen na hrvatski jezik. Uvezivanje audio datoteke odrađeno je kao i kod prethodna dva algoritma korištenjem biblioteke SpeechRecognition.

Odlomak koda 3 - Google Cloud Speech-to-Text

```
os.environ["GOOGLE_APPLICATION_CREDENTIALS"]="ključ.json"
client = speech.SpeechClient()
def transcribe_audio(audio_path):
    with open(audio_path, 'rb') as audio_file:
        content = audio_file.read()
     audio = types.RecognitionAudio(content=content)
     config = types.RecognitionConfig(
     encoding=enums.RecognitionConfig.AudioEncoding.LINEAR16,
       sample_rate_hertz=16000,
       language_code='hr-HR'
    )
    response = client.recognize(config, audio)
    for result in response.results:
        print('Text: ', result.alternatives[0].transcript)
```

8.4. Rezultati testiranja

Nakon pokretanja opisanih Python programa, svaki je od njih rezultirao svojim izlazom, tj. izlaznim tekstom koji su svi ispisani u tablici 1 za opći pregled rezultata. Na početku tablice nalazi se originalni ulomak te su u nastavku izlazni tekstovi podijeljeni na tekstove dobivene prepoznavanjem govora snimke ženskog glasa i prepoznavanjem govora snimke muškog glasa.

Tablica 1 - rezultati testiranja prepoznavanja govora

Originalni ulomak	No, kad se približiš, vidiš: to je velika bijela kuća. U njoj je škola. Kraj škole dvorište, u njemu djeca, đaka više od stotine. Sad možeš točno razabrati riječi. Djeca viču: "Stoj! Ne daj! K meni! Bježi! Trgaj! Pucaj! Skači! Hvataj ga! Drži, drž!" Jedni bježe dvorištem. Gaze jedan drugoga. Neki se parovi rvu. Ima ih koji jašu jedan na drugome. Većinom to jači i teži jaše na slabijemu. Neki, opet, skaču preko školskih drva. Ima ih koji preskakuju plotove. Neki se tuku. (Lovrak, 2003)	
	Ženski glas	*Muški glas*

Google Speech Recognition	no kad se približiš vidiš to je velika bijela kuća u njoj škola škola dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi dijeca viču stoji ne daj meni bježi trg aj pucaj kači hvataj ga drži drži jedni bježe dvorište gaze jedan drugoga neki se parovi robu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem u neki opet skaču preko školskih darova ima ih koji preskaču ju plotove neki se tuku	kad se približiš vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu dijeca đaka više od 100 sad možeš točno razabrati riječi i saviću stoji ne daj meni bježi trg aj pucaj skači hvataj ga drži teže jedni bježe dvorište gaze jedan drugoga neki se parovi rvo ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem neki opet skaču preko školskih treba ima ih koji preskaču plotove neki se tuku
Wit.ai	no kad se približi vidiš to je velika bijela kuća. u njoj je škola kroz škole dvorište u njemu djeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj k meni bježi trka i pucaj skače hvataju ga drži drž jedni bježe dvorištem gaze jedan drugoga neki se parovi zovu ima ih koji jašu jedan drugome većinom to jači i teži jaše na slabijem a neki opet skaču preko školskih darova ima ih koji prska kuju plodove neki se tuku	no kad se približi vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu djeca đaka više od stotina sad možeš točno razabrati riječi djeca viču stoji ne daj k meni bježi trgaju pucaju skače hvataju ga drži dere jedni bježe dvorištem gaze jedan drugoga neki se parovi zovu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem a neki opet skaču preko školskih trava ima ih koji preskaču plodove neki se tuku

Google Speech-to-Text Cloud Services	no kad se približiš vidiš to je velika bijela kuća u njoj škola škola dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoji ne daj meni bježi trg aj pucaj kači hvataj ga drži drži jedni bježe dvorište gaze jedan drugoga neki se parovi ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem u neki opet skaču preko školski hitlerova ima ih koji preskaču ju plotove neki se tuku	no kad se približiš vidiš to je velika bijela kuća kraj škole dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi deca biću stoji ne daj meni bježi trg aj pucaj skači hvataj ga drži bježe jedni bježe dvorište kaže jedan drugoga neki sve parovi ima ih koja su jedan na drugome većina jači se na slabije neki opet skaču prva ima ih koji preskaču plotove neki svetu

S obzirom na to da cilj testiranja nije bio „uhvatiti“ i interpunkcijske znakove, u nastavku će biti analizirani rezultati svakog algoritma zasebno, a originalni odlomak bit će prikazan bez interpunkcijskih znakova i malim slovima.

Originalni odlomak ima 83 riječi. Prepoznavanje govora prvim algoritmom prikazano je u tablici 2 te su točno prepoznate riječi označene zelenom bojom. Postoje razlike u izlaznim tekstovima za muški i ženski glas što potvrđuje hipotezu da algoritam prepoznavanja govora i pretvorbe govora u tekst neće raditi jednako za muški i ženski glas zbog razlika u visini tona glasova i njihovih frekvencija. Za odlomak izgovoren ženskim glasom prepoznato je 83 riječi od kojih je 68 riječi prepoznato točno. Za odlomak izgovoren muškim glasom prepoznate su 82 riječi od kojih je 69 riječi prepoznato točno. U oba slučaja postoje riječi koje nisu dobro prepoznate te se zato razlikuje ukupni zbroj riječi u izlaznim podacima. Umjesto riječi „preskakuju“ u oba slučaja je prepoznato „preskaču“, što je slučajno i glagol sličnog značenja, a riječ „djeca“ prepoznata je kao „dijeca“ što nije riječ koja uopće postoji u hrvatskom jeziku. Postoje i neke riječi koje su prepoznate kao dvije, primjerice umjesto „trgaj“ prepoznato je „trg aj“. U odlomku izgovorenom ženskim glasom nije prepoznata riječ „kraj“ ispred riječi „škole“, a nakon riječi „preskaču“ prepoznata je i riječ „ju“ pa je „preskakuju“ postalo „preskaču ju“. Kratica „drž“ ni u jednom slučaju nije točno prepoznata, već je sa snimke ženskog glasa prepoznata kao „drži“, a sa snimke muškog glasa je prepoznata kao „teže“. Zanimljivo je primijetiti da je u oba slučaja riječ „dvorištem“ prepoznata kao „dvorište“, tj. da nije prepoznat suglasnik „m“ na kraju riječi.

Tablica 2 - rezultati prepoznavanja govora za Google Speech Recognition

Originalni odlomak	*Ženski glas*	*Muški glas*
no kad se približiš vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu djeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj k meni bježi trgaj pucaj skači hvataj ga drži drž jedni bježe dvorištem gaze jedan drugoga neki se parovi rvu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijemu neki opet skaču preko školskih drva ima ih koji preskakuju plotove neki se tuku	no kad se približiš vidiš to je velika bijela kuća u njoj škola škola dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi dijeca viču stoji ne daj meni bježi trg aj pucaj kači hvataj ga drži drži jedni bježe dvorište gaze jedan drugoga neki se parovi robu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem u neki opet skaču preko školskih darova ima ih koji preskaču ju plotove neki se tuku	kad se približiš vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu dijeca đaka više od 100 sad možeš točno razabrati riječi i saviću stoji ne daj meni bježi trg aj pucaj skači hvataj ga drži teže jedni bježe dvorište gaze jedan drugoga neki se parovi rvo ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem neki opet skaču preko školskih treba ima ih koji preskaču plotove neki se tuku

Prepoznavanje govora uz pomoć alata Wit.ai donio je nešto bolje rezultate od Googleovog običnog Speech Recognition algoritma. Prepoznavanje govora za Wit.ai prikazano je u tablici 3 te su opet zelenom bojom označene riječi koje su točno prepoznate. Za odlomak izgovoren ženskim glasom prepoznato je 85 riječi od kojih je 71 riječ prepoznata točno. Za odlomak izgovoren muškim glasom prepoznate su 84 riječi od kojih je 70 riječi prepoznato točno. Pretvorba govora u tekst i ovim algoritmom ima neke nedostatke koji se pojavljuju u oba slučaja. Primjerice, riječ „približiš“ u oba je slučaja prepoznata kao „približi“, tj. bez suglasnika „š“ na kraju riječi te je riječ „plotove“ u oba slučaja prepoznata kao „plodove“. Riječ „trgaj“ opet nije dobro prepoznata u oba slučaja, no u pretvorbi ženskog govora u tekst prepoznata je kao dvije riječi „trka i“, a u pretvorbi muškog govora u tekst prepoznata je kao riječ „trgaju“. Također, mnogi imperativi originalnog odlomka koji su u drugom licu jednine („trag“, „pucaj“, „skači“, „hvataj ga“) iz audio snimke muškog glasa prepoznati su kao treće lice prezenta („trgaju“, „pucaju“, „skače“, „hvataju ga“). Riječ „drva“ kao ni kod prepoznavanja prvim algoritmom nije točno prepoznata te je u slučaju prepoznavanja govora alatom Wit.ai prepoznata kao „darova“ kod prepoznavanja ženskog glasa, a „trava“ kod prepoznavanja muškim glasom. Riječi su slične jer započinju sličnim suglasnicima, sadrže suglasnike „r“ i „v“ te samoglasnik „a“ što je naznaka prepoznavanja fonema i takvog sastavljanja riječi. Riječ „drž“ je samo u slučaju prepoznavanja govora ženskog glasa ovim alatom točno prepoznata.

Tablica 3 - rezultati prepoznavanja govora za Wit.ai

Originalni odlomak	*Ženski glas*	*Muški glas*
no kad se približiš vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu djeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj k meni bježi trgaj pucaj skači hvataj ga drži drž jedni bježe dvorištem gaze jedan drugoga neki se parovi rvu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijemu neki opet skaču preko školskih drva ima ih koji preskakuju plotove neki se tuku	no kad se približi vidiš to je velika bijela kuća u njoj je škola kroz škole dvorište u njemu djeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj k meni bježi trka i pucaj skače hvataju ga drži drž jedni bježe dvorištem gaze jedan drugoga neki se parovi zovu ima ih koji jašu jedan drugome većinom to jači i teži jaše na slabijem a neki opet skaču preko školskih darova ima ih koji prska kuju plodove neki se tuku	no kad se približi vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu djeca đaka više od stotina sad možeš točno razabrati riječi djeca viču stoji ne daj k meni bježi trgaju pucaju skače hvataju ga drži dere jedni bježe dvorištem gaze jedan drugoga neki se parovi zovu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem a neki opet skaču preko školskih trava ima ih koji preskaču plodove neki se tuku

Prepoznavanje govora Google Cloudovim Speech-to-Text API-jem je s obzirom na najveću količinu potrebnog uloženog truda i najveću komercijalnu cijenu alata donijelo poražavajuće rezultate. Alat je samo malo bolji od besplatne verzije Googleovog Speech Recognizera. Prepoznavanje govora Google Cloud Speech-to-Text API-jem prikazano je u tablici 4, a sve točno prepoznate riječi označene su zelenom bojom. Za odlomak izgovoren ženskim glasom prepoznata je 81 riječ od kojih je 69 riječi prepoznato točno. Za odlomak izgovoren muškim glasom prepoznato je 79 riječi od kojih je 70 riječi prepoznato točno. Ponavljaju se neke greške pri prepoznavanju govora koje su se pojavile i u testiranju prvog algoritma, primjerice prepoznavanje riječi „djeca“ kao riječ „dijeca“ koja nije hrvatska riječ te u prepoznavanju ženskog glasa nije prepoznata riječ „kraj“ prije riječi „škole“. Također je opet u oba slučaja riječ „dvorištem“ prepoznata kao „dvorište“, tj. nije prepoznat suglasnik „m“ na kraju riječi, a osim toga ovaj put nema na kraju riječi „školskih“ suglasnika „h“ već je riječ prepoznata kao „školski“. Kod prepoznavanja govora u audio snimci ženskog glasa riječ „slabijemu“ prepoznata je kao dvije riječi „slabijem u“, dok je kod muškog glasa prepoznata točno. Prepoznavanje govora ženskog glasa riječ „skači“ prepoznalo je kao riječ „kači“ te je to jedini slučaj u cijelom testiranju da nije prepoznat suglasnik s početka riječi (za slučaj da je riječ uopće prepoznata).

Tablica 4 - rezultati prepoznavanja govora za Google Cloud Speech-to-Text API

Originalni odlomak	*Ženski glas*	*Muški glas*
no kad se približiš vidiš to je velika bijela kuća u njoj je škola kraj škole dvorište u njemu djeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj k meni bježi trgaj pucaj skači hvataj ga drži drž jedni bježe dvorištem gaze jedan drugoga neki se parovi rvu ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijemu neki opet skaču preko školskih drva ima ih koji preskakuju plotove neki se tuku	no kad se približiš vidiš to je velika bijela kuća u njoj škola škola dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoji ne daj meni bježi trgaj pucaj kači hvataj ga drži drži jedni bježe dvorište gaze jedan drugoga neki se parovi ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijem u neki opet skaču preko školski hitlerova ima ih koji preskaču ju plotove neki se tuku	no kad se približiš vidiš to je velika bijela kuća u njoj škola kraj škole dvorište u njemu dijeca đaka više od stotine sad možeš točno razabrati riječi djeca viču stoj ne daj meni bježi trgaj pucaj skači hvataj ga drži brže jedni bježe dvorište kaže jedan drugoga neki sve parovi ima ih koji jašu jedan na drugome većinom to jači i teži jaše na slabijemu neki opet skaču prva školski ima ih koji preskaču plotove neki sve tuku

Bez obzira na greške u prepoznavanju govora, sva tri alata pokazala su visok stupanj točnosti prepoznatih riječi. Prosječan broj točno prepoznatih riječi je 69,5 riječi. U tablici 5 prikazan je pregled broja prepoznatih riječi po alatima i prema razlici u muškom i ženskom glasu. Točnost prepoznatih riječi izračunata je podjelom broja točno prepoznatih riječi s ukupnim brojem riječi iz originalnog odlomka. Prosječna točnost prepoznavanja govora svih testiranih alata je 83,7%.

Google Speech Recognition u prosjeku je prepoznao 68,5 riječi i imao postotak od 82,5% za pretvorbu govora u tekst za muški i ženski glas. Wit.ai je alat s kojim je u prosjeku prepoznato 70,5 riječi s postotkom točnosti 84,9% u prosjeku. Google Cloud Speech-to-Text API u prosjeku je prepoznao 69,5 riječi i imao prosjek točnosti 83,7% za snimke muškog i ženskog glasa.

U prosjeku su svi alati malo bolje prepoznavali snimke muškog glasa s prosječnih 69,6 riječi što je postotak točnosti 83,9% dok su snimke ženskog glasa imale prosječno 69,3 točno prepoznatih riječi, tj. postotak od 83,5%. Bez obzira na to, najbolje rezultate pokazao je alat Wit.ai na uzorku ženskog glasa sa 71 prepoznatom riječi i postotkom točnosti 85,5%.

Tablica 5 - pregled prepoznatih riječi po alatima

alat		Broj prepoznatih riječi	Broj točno prepoznatih riječi	Točnost (broj prepoznatih riječi/broj riječi)
Google Speech Recognition	Ženski glas	83	68	0,819
	Muški glas	82	69	0,831
Wit.ai	Ženski glas	85	71	0,855
	Muški glas	84	70	0,843
Google Cloud Speech-to-Text API	Ženski glas	81	69	0,831
	Muški glas	79	70	0,843

9. Zaključak

U ovom je radu istraženo područje tehnologije pretvorbe govora u tekst i prepoznavanja govora te njezina primjena za hrvatski jezik uz pomoć alata dostupnih pomoću programskog jezika Python. Kroz ovo istraživanje stečen je uvid u napredak, izazove i potencijal ove domene.

Predstavljeno testiranje postojećih algoritama za pretvorbu govora u tekst koje je moguće primijeniti na hrvatski jezik pokazalo je učinkovitost Pythona kao korisnog jezika za implementaciju algoritama pretvorbe govora u tekst. Dostupnost različitih biblioteka i okvira u Pythonu, kao što je SpeechRecognition, omogućuju postavljanje temelja i daljnji razvoj učinkovitih i točnih modela za pretvaranje govornog hrvatskog jezika u pisani tekst.

Pri razvoju algoritama i tehnologije za pretvorbu govora u jezik i prepoznavanje govora važno je promatrati karakteristike i nijanse specifične za svaki jezik pri dizajniranju sustava za prepoznavanje govora. Hrvatski je jezik sa svojom jedinstvenom fonetskom i morfološkom strukturom te bogatim vokabularom predstavio specifične izazove koji su zahtijevali pažljivo razmatranje tijekom razvoja algoritma.

Zaključno, ovaj rad i testiranje provedeno u sklopu rada naglašavaju značajne pomake u tehnologiji pretvorbe govora u tekst i njezinoj primjeni na hrvatskom jeziku. Upotrebom programskog jezika Python i biblioteke SpeechRecognition prikazana je mogućnost izvedbe takvog prepoznavanja govora za hrvatski jezik te njen potencijal točnog i učinkovitog prepoznavanja govora. Kako se područje nastavlja razvijati rastućim brojem podataka, tehnologija pretvorbe govora u tekst mogla bi igrati ključnu ulogu u načinu na koji komuniciramo govornim jezikom što koristi raznim industrijama, a u slučaju hrvatskog jezika, i njegovoj većoj dostupnosti. Uspješna implementacija algoritama za pretvaranje govora u tekst na hrvatskom jeziku otvara nove mogućnosti za poboljšanje komunikacije, pristupačnosti jezika i učinkovitosti u različitim sektorima, uključujući obrazovanje, zdravstvo, službu za korisnike i tako dalje. Pridavanjem važnosti interdisciplinarnoj suradnji, lingvističkom proučavanju i algoritamskom usavršavanju, postavljaju se temelji za budući razvoj na području tehnologije pretvorbe govora u tekst za hrvatski, ali i sve druge jezike.

Literatura

1. Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access, vol. 10*, str. 122136-122158.
2. Chou, W., & Juang, B. (2003). *Pattern Recognition in Speech and Language Processing.* Boca Raton: CRC Press.
3. CMUSphinx. (2023). *Open Source Speech Recognition Toolkit*. Preuzeto 26. svibnja 2023 iz CMUSphinx: https://cmusphinx.github.io/
4. Google. (2023). *Google Speech-toText*. Preuzeto 26. svibnja 2023 iz Google Cloud: https://cloud.google.com/speech-to-text/
5. Hrvatska enciklopedija. (2021). *Hrvatska enciklopedija, mrežno izdanje*. Preuzeto 16. svibnja 2023 iz https://enciklopedija.hr
6. HTK. (2016). *HTK3*. Dohvaćeno iz https://htk.eng.cam.ac.uk/
7. Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* New Jersey: Prentice Hall.
8. InternationalPhoneticAlphabet.org. (2015). *IPA International Phonetic Alphabet.* Preuzeto 18. svibnja 2023 iz https://www.internationalphoneticalphabet.org/wp-content/uploads/2016/08/IPA-Chart-Kiel-Font-2015.pdf
9. Ipšić, I., & Martinčić-Ipšić, S. (2010). Croatian Speech Recognition. U N. Shabtai, *Advances in Speech Recognition* (str. 123-140). London: Sciyo.
10. Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* New Jersey: Prentice Hall.
11. Kaldi. (2023). *Kaldi*. Preuzeto 26. svibnja 2023 iz Kaldi: https://www.kaldi-asr.org/doc/about.html
12. Lovrak, M. (2003). *Vlak u snijegu.* Zagreb: Mozaik knjiga.
13. Microsoft Azure. (2023). *Speech*. Preuzeto 26. svibnja 2023 iz Azure: https://azure.microsoft.com/en-us/products/cognitive-services/speech-services/
14. Mozilla Corporation. (2023). *Mozilla DeepSpeech*. Preuzeto 26. svibnja 2023 iz Mozilla DeepSpeech: https://deepspeech.readthedocs.io/en/r0.9/#

15. Pisanski, K. (2014). *Human vocal communication of body size.* ResearchGate.

16. Python Software Foundation. (2023). *Python*. Dohvaćeno iz https://www.python.org/

17. Python Software Foundation. (2023). *SpeechRecognition*. Dohvaćeno iz https://pypi.org/project/SpeechRecognition/

18. Rabiner, L. R., & Schafer, R. W. (2007). Introduction to Digital Speech Processing. *Foundations and Trends in Signal Processing*, str. 1-194.

19. Šoić, R., & Vuković, M. (ožujak 2022). N-gram Based Croatian Language Network: Application in a Smart Environment. *Journal of Communications Software and Systems, vol.18*, str. 63-71.

20. Tadić, M., Brozović-Rončević, D., & Kapetanović, A. (2012). *Hrvatski jezik u digitalnom dobu.* Heidelberg: Springer.

21. Weaver, W., & Shannon, C. E. (1964). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

22. wit.ai. (2023). *wit.ai*. Preuzeto 26. svibnja 2023 iz wit.ai: https://wit.ai/

23. Yu, D., & Deng, L. (2015). *Automatic Speech Recognition.* London: Springer.

24. Zulkifly, M., & Yahya, N. (2017). Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD). *2017 IEEE 3rd International Symposium in Robotics and Manufacturing Automation (ROMA)*, (str. 1-5). Kuala Lumpur.

Popis slika

Popis tablica

Popis odlomaka koda

Govorne tehnologije u programskom jeziku Python za hrvatski jezik

Sažetak

Govorne tehnologije, posebno govor-u-tekst (engl. *speech-to-text*), zahvaljujući razvitku umjetne inteligencije i posebno neuronskih mreža iznimno su napredovale posljednjih godina i postale ključnim alatom u mnogim područjima, kao i u svakodnevici.

Cilj ovog diplomskog rada je opisati kako funkcionira tehnologija pretvaranja govora u tekst. Rad se sastoji od dva dijela. U prvom, teorijskom dijelu bit će navedeni i objašnjeni neki od ključnih izazova s kojima se susrećemo pri stvaranju sustava pretvorbe govora u tekst. Također će se navesti ograničenja i izazovi koji su povezani s tehnologijom pretvaranja govora u tekst te će se opisati razni alati koje je moguće koristiti za pretvorbu govora u tekst. Drugi, praktični dio rada bavit će se korištenjem algoritama za pretvorbu govora u tekst. U radu će se napraviti usporedna analiza i testiranje algoritama u programskom jeziku Python za hrvatski jezik.

Ključne riječi: govor u tekst, govorne tehnologije, Python

Speech technologies in the Python programming language for the Croatian language

Summary

Speech technologies, especially speech-to-text, thanks to the development of artificial intelligence and especially neural networks, have progressed tremendously in recent years and have become a key tool in many areas, as well as in everyday life.

The aim of this master's thesis is to describe how speech-to-text technology works. The paper consists of two parts. In the first, theoretical part, some of the key challenges we face when creating a speech-to-text conversion system will be listed and explained. It will also outline the limitations and challenges associated with speech-to-text technology and describe the various tools that can be used for speech-to-text. The second, practical part of the paper will deal with the use of algorithms for converting speech into text. In the paper, a comparative analysis and testing of algorithms in the programming language Python for the Croatian language will be done.

Key words: speech-to-text, speech technologies, Python