Identifikacija govornika iz snimki glasa

Lipovac, Antonio

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Antonio Lipovac

Identifikacija govornika iz snimki glasa

ZAVRŠNI RAD

Varaždin, 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Antonio Lipovac

Matični broj: 0016142555

Studij: Informacijski sustavi

Identifikacija govornika iz snimki glasa

ZAVRŠNI RAD

Mentor/Mentorica:

Doc. dr. sc. Petra Grd

Varaždin, kolovoz 2022

Antonio Lipovac

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

Završni rad na temu Identifikacija govornika iz snimki glasa istražuje što je to zapravo identifikacija govornika, za što se sve može koristiti i gdje je se prvo pojavila. Kao što druge identifikacije spadaju u Biometriju tako spada i identifikacija govornika. Kroz rad prikazano je kako su ljudi od početka pa do danas primjenjivali i na koji način su implementirali identifikaciju govornika. Nakon toga objašnjen objašnjeno je što je to zvuk i na koji način se obrađuje te stvara njegov digitalni oblik. Objašnjeno je detaljno što zapravo je identifikacija govornika i koja je razlika između identifikacije govornika i govora, te su prikazane glavne značajke prepoznavanja govornika. Na kraju se objašnjava moderno korištenje prepoznavanja govornika pomoću dubokog učenja.

Ključne riječi: prepoznavanje/ identifikacija govornika, izjava, duboko učenje

Sadržaj

1. Uvod

Danas ljudi sve više koriste neke oblike biometrije, a da toga nisu ni svjesni. Praktički svatko ima otisak prsta na mobitelu ili čak prepoznavanje lica što su dva najpoznatija biometrijska sustava baš iz razloga što se svi svakodnevno susreću s njima. Ali osim njih sve više se koristi prepoznavanje govornika ili identifikacija govornika koju ljudi često miješaju s prepoznavanjem govora. Dok nam prepoznavanje govora daje odgovor što je osoba izgovorila, prepoznavanje govornika nam daje odgovor tko je osoba koja govori. Neki od najpoznatijih sustava identifikacije govornika su Googleov sustav i Amazonov sustav koji prepoznaju svoje korisnike i mogu točno za određenog korisnika pamtiti konkretno njegove bitne stvari kao što su datumi, neke činjenice i druge korisniku važne stvari.

Nekada su ljudi vršili prepoznavanje govornika tako što su obrazovali ljude koji bi imali posao samo razlikovati glasove, ali kako svaki čovjek ima drukčiju percepciju i raspoloženje brzo je taj način zaboravljen i stvoreni su algoritmi koji su se bavili prepoznavanjem govornika. Od tog dana do danas ti se algoritmi sve više razvijaju i postaju točniji, jeftini i brži, a tome je pomoglo razvijanje dubokog učenja koje je još više olakšalo identifikaciju govornika.

2. Biometrija

Biometrija kao pojam potječe od grčkih riječi bios što predstavlja život i metron što predstavlja mjeru, pa tako biometrija predstavlja mjerenje određenih tjelesnih i ponašajnih karakteristika živih bića. Tehnologija koja analizira i mjeri ponašajne i fizičke karakteristike čovjeka naziva se biometrijska identifikacija [3]. Glavna predispozicija i namjena biometrije je sama identifikacija korisnika na razne načine, ali ne mora se nužno koristiti samo za identifikaciju već se može koristiti i u drugim područjima kao što je prepoznavanje glasa govornika u svrhu pretraživanja preko Google tražilice. Jedan od elemenata biometrije je pretvaranje analognog signala u digitalni i pritom stvaranje umjetne inteligencije kako bi računalo moglo samostalno pamtiti i koristiti unesene podatke to jest informacije [6].

Biometrijske karakteristike se dijele na fizičku biometriju i biometriju ponašanja. Fizička biometrija kako joj i samo ime govori bazira se na fizičkoj posebnosti to jest jedinstvenosti osobe. Kako je poznato da je svaka osoba jedinstvena tako se fizičke predispozicije mogu koristit kako bi se određena osoba identificirala. U tu metodu spadaju: prepoznavanje crta lica, raspored vena, geometrija dlana, DNK zapis, skeniranje oka i otisak prsta. S druge strane biometrija ponašanja bavi se ponašajnim karakteristikama čovjekova tijela koja su također jedinstvena za svaku osobu. Informacije koje se dobivaju ovim modelom većinom su prikazane krivuljama koje se koriste za opis ponašanja te se iste te krivulje koriste za identifikaciju. U model biometrijskih ponašanja spadaju: prepoznavanje glasa, rukopisa, tipkanja te tjelesni mirisi [6].

2.1. Prepoznavanje glasa kao biometrijska karakteristika

Svaki čovjek ima jedinstven glas isto kao što ima jedinstven i otisak prsta [2]. To se vrlo jednostavno može zaključiti samo po razgovoru s nekoliko ljudi. Istina je da se može naći ljudi koji imaju jako slične glasove, ali svaki čovjek ipak ima jedinstven glas ista stvar je s izgledom. Ljudi mogu izgledati slično ali ne u potpunosti isto. Neke od glavnih stvari koje stvaraju glas jedinstvenim su jezik kojim osoba govori ili možda čak govori više jezika, naglasak koji koristi i uz koji je odrastao te dijalekt koji koristit. Također veliku ulogu ima veličina i oblik govornih organa osobe koji se sastoje od vokalnih nabora i vokalnog trakta. Pa na primjer muškarci imaju dublji, žene imaju viši glas, dok djeca imaju nježan [1].

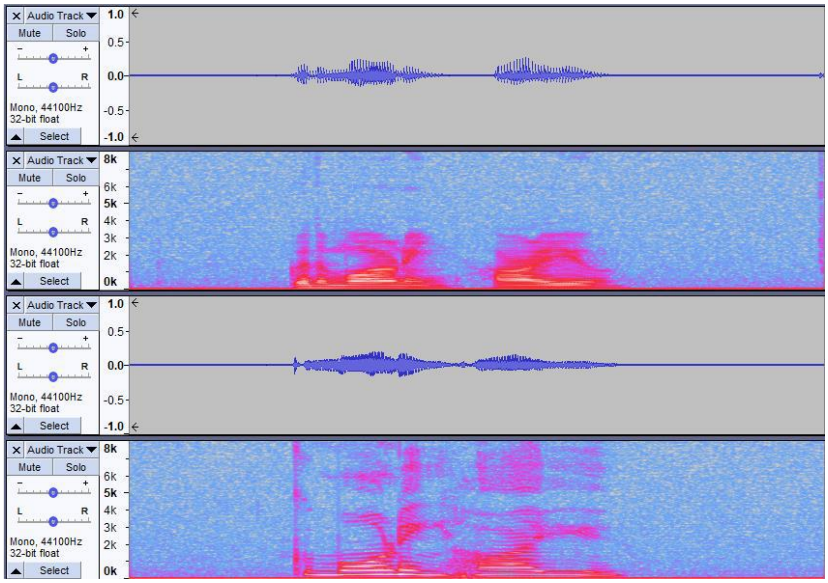
Područje identifikacije glasa zove se fonoskopska identifikacija, te se ona bavi određenim karakteristikama glasa kao što su frekvencija, boja glasa, govorne mane i slično. Sam postupak identifikacije se mijenjao kroz povijest te je bio određivan na razne načine, ali

sve je bazirano na pretvaranju glasa u digitalni oblik koji se pomoću raznih algoritama može podijeliti u dijelove te nakon toga uspoređivati i moći donositi odluke koliko je zapravo koji glas sličan ili različit od drugoga. Sinonimi za prepoznavanje glasa govornika (speaker recognition) su identitet govornika (speaker identity), id-glasa (voice-id) ili id-govornika (speaker-id) i glasovni otisak (voiceprint), ali najčešće korišten je speaker recognition to jest prepoznavanje glasa govornika [4]. Prepoznavanjem govornika se mogu baviti razni stručnjaci. Većina njih se bavi fonetikom, lingvistikom, patologijom glasa i govora ili akustikom [7].

3. Povijest prepoznavanja glasa i tehnike

3.1. Najranije tehnike prepoznavanje govornika

Prvi zapisani podatak da je policija koristila prepoznavanje govornika dolazi iz 1927. godine iz Circleville Herald-a iz SAD-a gdje je se spremao glas svih zatvorenika. Nakon toga prvi pisani podatak o prepoznavanju govornika dolazi nam iz 1935 sa konferencije u Pittsburgu gdje se govori kako je identifikacija govora bolja od identifikacije rukopisa, te kako je lakša za potvrditi nacionalnost. Te se za te izvore koristio oscilograf. Prvi izvor koji je koristio spektogram bilo je akademsko istraživanje 1962. iz koje je i nastala tvrtka Voiceprint Laboratories koja je svoj rad bazirala na ljudskom čitanju i imala je 97% točnosti prilikom prepoznavanja govornika. Oni su radili istraživanje na 15 muškaraca i 10 žena na način da su izabrali 10 visokofrekventnih riječi koje su čitali u različitim kontekstima te su kasnije uspoređivali njihove spektograme. Sljedeća slika prikazuje usporedbu izgleda oscilografa i spektograma, te se može primijetiti kako nam spektogram daje puno korisniju sliku po kojoj bi mogli vršiti usporedbe [4].



Slika 1 Usporedba oscilografa i spektograma (Izvor: izrada autora prema Wang, Q. 2022.)

3.2. Počeci razvoja algoritama

Algoritmi su se počele razvijati zbog glavnog problema koji je zapravo bio u ljudima. To jest ljudi su bili oni koji su se bavili otkrivanjem koji glas kome pripada i uspoređivali su glasove ljudi. Tu su se stvorila dva problema, jedan je bio cijena jer su se ljudi morali plaćati i obrazovati a isto tako moralo se ljude zapošljavati na puno mjesta. A drugi problem bio je to što nisu svi ljudi isti te je nemoguće da svi jednako dobro odrađuju taj posao. Iz toga razloga morali su se razviti algoritmi za prepoznavanje govornika. Oni su bili skupi za razviti ali kasnije su se mogli prilagoditi i koristiti na više mjesta, te su bili od povjerenja jer su uvijek za iste unose davali iste izlaze [4].

Najraniji algoritam bazirao se na podudaranju uzoraka i bio je inspiriran ljudskim čitanjem. Umjesto ručne usporedbe, on je matematički uspoređivao sadržaj, ali sadržaj je morao biti isti. Također algoritam je imao manu da nije uzimao u obzir moguće varijacije kao što su različiti uređaji za snimanje, okolna buka ili čak emocije. Način da se taj problem riješi bio je da se radi statistika to jest da se za svakog govornika više puta snima glas na razne načine i na raznim mjestima. Za poboljšanje su, osim srednje vrijednosti, uzimali i standardnu devijaciju te su koristili statistiku predsegmenta. Najčešće korišteni statistički model prije deep learninga bio je Gaussian Mixture Models (GMM). Njegova prednost je bila što je jednostavan model. On je kompleksne podjele zbrajao u jednostavne Gaussianove podjele. Postoje dvije varijante: Gaussian Mixture Models-univerzalni pozadinski model (GMM-UBM) i Gaussian Mixture Models- support vector machines (GMM–SVM). Model se koristio do 2006. Godine [4].

3.3. Razvoj modernih algoritama

Umjetna inteligencija postaje dominantan pristup oko 2010. te se sve više koristi za prepoznavanje govora i govornika, za računalni vid i za procesiranje govora (natural language processing NLP). Prednost je bila ta što koristi jednostavne funkcije, to jest neurone, za pojednostavljenje složenih funkcija. Prepoznavanje govornika počinje se bazirati na modelu ponavljajućih neuronskih mreža (recurrent neural networks - RNN) što je dovelo do toga da se neke od organizacija istaknu pomoću svojih algoritama i sustava unutar umjetne inteligencije. Prvi je bio Google s d-vektor modelom, napravili su prvi sustav prepoznavanja govornika pomoću deep neural network-a (DNN), prvi end-to-end sustav i sustav koji se bazira na pažnji, zatim Baidu sa sustavom za prepoznavanju govornika, Sveučilište Johns Hopkins (JHU) sustavom x-vectora i Sveučilište Shanghai Jiao Tong (SJTU) sa sustavom j-vektora. Razvijeni su i prvi pametni asistenti, a prvi je to napravio iPhone s iPhone 4S mobilnim uređajem koji je imao Siri glasovnu pomoćnicu. Nakon toga Amazon razvija Amazon Echo prvi pametni kućni

uređaj i nakon njih se otvaraju vrata i drugim organizacijama, te se sve više stvara potreba za prepoznavanjem govornika u svrhu sigurnosti ili personalizacije [4].

4. Temeljna obrada zvuka

4.1. Audio i akustika

Kako se želi konkretno definirati prepoznavanje govornika to se ne može napraviti bez da se prvo definira zvuk. Najbitniji audio signal za prepoznavanje govornika je govor što zapravo predstavlja sve zvukove koje artikuliraju ljudi koji su društveno smisleni, a njime se obično izražavaju misli i osjećaji [4]. Utterance ili na hrvatski iskaz, izjava ili izgovaranje vrlo je bitna riječ u prepoznavanju govornika, te će biti kroz rad često korištena. Iskaz se definira kao audioisječak govornih signala, a ovisi o govorniku, jeziku govornika i frazama koje govornik izgovara [4]. Za zvuk je opće poznato da je to vibracija koja se širi kao mehanički longitudinalni val te može prolaziti kroz sva agregacijska stanja [10]. Sam govor stvaraju vokalni organi te svaki dio organa ima svoju ulogu, pa tako vokalni nabori stvaraju vibracije, glasnice kontroliraju protok zraka, a vokalni trakt modulira zvuk. Zvučni valovi na slici mogu izgledati vrlo kompliciranо ali većinom se radi o više sinusoida koje se nastavljaju jedna na drugu [10], [11].

4.2. Sluh i percepcija

Poznato je da je glavna karakteristika zvuka frekvencija, a osim nje mjeri se i amplituda. Oboje su fizičke veličine koje se ne mogu intuitivno zaključiti samo iz perspektive ljudskog uha nego postoje uređaji koji ih mjere. Ljudi mogu čuti frekvencije od 20Hz do 20 kHz, sve ispod 20Hz naziva se infrazvuk dok sve više od 20kHz naziva se ultrazvuk [10], [12]. Potrebno je znati i za pojam rezonanca koji se dešava kada se primjeni sila jedna prirodnoj frekvenciji, prilikom tog dešavanja povećava se amplituda i stvara stojni val koji povećavaju glasnoću zvuka. Za zvuk je potrebno znati i što je intenzitet zvuka. Intenzitet zvuka je zapravo količina energije koja prođe kroz određenu površinu u određenom vremenu, a još se naziva i jakost te se mjeri u decibelima (dB) [12].

Poznato je da je ljudska percepcija zvuka nelinearna jer je određena fiziološkom strukturom. Pa tako postoje dva aspekta koja utječu na nelinearnost a to su: nelinearnost frekvencije i nelinearnost intenziteta. Postoje dvije ljestvice po kojima je definirana nelinearnost frekvencije. Prva je Barkova ljestvica, to je psihoakustička ljestvica koju je definirao Ederhard Zwicker. Namjena joj je bila da jednaka udaljenost na ljestvici odgovara jednakoj visini frekvencije. Pa je tako svaka frekvencija niža od 500Hz na grafu prikazana kao linearna funkcija, a iznad kao logaritamska. Druga ljestvica je Melova ljestvica, te je ona vrlo slična Barkovoj, ali je za razliku od Barkove ona kontinuirana te uvodi mjernu jedinicu za melodiju Mel. 1000Hz je približno 1000mel po Melovoj ljestvici. Najveći problem nelinearnosti intenziteta

je ljudski sluh. Ljudski sluh je osjetljiviji na signale s manjom amplitudom, na primjer može čuti kako igla pada u tihom okruženju, ali je manje osjetljiv na signale s velikom amplitudom. Jedini način da se to stimulira je korištenje logaritamske funkcije i kubnog korijena [4].

4.3. Obrada audio signala

Nije moguće izravno izmjeriti zvučni tlak niti neutralne električne signale. Potrebni su drugi načini prikupljanja, pohranjivanja i analize govornih signala. Signali su predstavljeni u dva oblika, a to su analogni i digitalni signali. Glavna naprava za pretvorbu signala je mikrofon. Mikrofon je sonda koja pretvara zvučne valove u električne signale i djeluje pomoću mehanizma za hvatanje vibracija zraka. Postoji nekoliko vrsta mikrofona kao što su: kondenzatorski, dinamički, MEMS mikrofon i mikrofonske vrpce. Kako bi koristili audio uređaje potreban nam je i audio poveznik, a taj poveznik dijelimo na muški utikač i ženski utičnica, također ga dijelimo na veličine, a najpoznatije su 6.35mm, 3.5mm koji je i najčešći i 2.5mm [4].

Računalo kada pohranjuje i obrađuje audio signale, oni su predstavljeni kao digitalni signali, pa tako svako računalo mora imati zvučnu karticu koja sadrži analogno-digitalni pretvarač (ADC) i digitalno-analogni pretvarač (DAC). Sama pretvorba ima određene zahtjeve kako bi se promjena desila, a oni su: uzorkovanje, kvantizacija i kodiranje. Analogno-digitalna pretvorba zapravo je kodiranje informacija prilikom koje se javlja određeni gubitak tih istih informacija. Sam gubitak zapravo je samo prilikom kvantizacije, dok kod kodiranja i uzorkovanja nema gubitaka. Uzorkovanje ili diskretizacija po vremenu i prostoru postoji iz dva razloga. Prvi razlog je da digitalni sustavi rade samo s konačnim i diskretnim skupovima podataka, a drugi razlog je da se uzorkovanjem ne gube podatci i umjesto da se radi na razini kontinuiranog signala radi se sa uzorcima pa se tako stvara ušteda prostora i sustavi koji međusobno sudjeluju mogu biti raspoređeni na više korisnika [16]. Prilikom uzimanja uzoraka smanjuje se signal neprekidnog vremena na signal diskretnog vremena. Bitno je prilikom uzimanja uzoraka znati što je stopa uzorkovanja ili brzina uzorkovanja (sampling rate), to je broj uzoraka snimljenih svake sekunde. S obzirom na brzinu uzorkovanja, možemo konstruirati samo periodične signale čija je frekvencija najviše upola manja od brzine uzorkovanja. Preciznost signala traje 4 bajta, a dvostruka preciznost traje 8 bajtova, zato kvantificiramo ove brojeve na cijele brojeve kako bi uštedili prostor, ali ćemo tako malo izgubiti na preciznosti [4]. Kvantitzacija ili diskretizacija po amplitudi je proces pretvaranja uzoraka iz kontinuiranog signala, a kako je skup kontinuiranih vrijednosti beskonačan mi koristimo kvantizaciju za podjelu tog skupa u intervale i prilikom podjele dešavaju se gubitci informacija [16]. Vrste kvantizacije su: 8-bitna ima 256 razina, 16-bitna 65536 razina i 24-bitna 16.8 milijuna razina [4].

4.4. Audio kodiranje i formati

Nakon uzimanja uzoraka i kvantizacije zvuk je predstavljen kao niz cijelih brojeva. Sam postupak kodiranja odvija se načinima uzorak po uzorak, bit po bit ili kvant po kvant. Za više prostora za pohranu i prijenos i pri nižim brzinama, zvuk može biti predstavljen u obliku bajtova, dok je uzorak po uzorak namijenjen većim brzinama. Najsporiji postupak je kodiranje kvant po kvant, ali isto tako radi se i o najjednostavnijem postupku [16]. Kodiranjem se dolazi od digitalnog signala do bajtova, a dekodiranjem od bajtova do digitalnog signala [4].

Audio formati određuju kako se bitovi koriste za kodiranje signala. Postoje različite vrste formata kao što su nekomprimirani, kompresija bez gubitaka i kompresija s gubitcima [4]. Nekomprimirani formati su snimke stvarnih valova u digitalnom obliku, pa se tako radi o najtočnijim i najvećim formatima. PCM ili modulacija pulsnog koda najčešći je audio format. Radi se o digitalnom prikazu neodređenih analognih signala, a on se dobije na način da se valni oblik analognog zvuka mora snimiti u određenim intervalima. PCM sadrži dubinu bita (eng. bit depth) i brzinu uzorkovanja (eng. sampling rate). Stvoren je i podtip PCM nazvan LPCM ili linearna modulacija pulsnog koda koji linearno prikuplja uzorke, te je on najpopularniji tip PCM-a danas. WAV (Waveform audio file format) ili format zvučne datoteke valnog oblika standard je koji su razvili IBM i Microsoft 1991. godine. WAV datoteka sadrži nekomprimirani zvuk i samo je omotač za PCM kodiranje. AIFF (Audio Interchange File Format) je kratica za format datoteke audio razmjene koju je razvio Apple za Mac sustave 1988. godine. Kao i WAV većinom sadrži nekomprimirani PCM format zvuka. I ako su WAV i AIFF stvoreni za određene sustave mogu se otvarati u svima bez problema [17].

Audio formati s kompresijom s gubicima stvoreni su iz razloga što su nekomprimirani zauzimali previše prostora. Točno je da su oni manje kvalitete jer se kompresijom gube određene značajke, ali zato zauzimaju manje prostora [13]. MP3 je najpopularniji audio format ovog tipa te je osnovan 1993. godine. Danas svaki uređaj podržava MP3 i može ga reproducirati. MP3 je kratica za MPEG-1 Audio Layer 3 i cilj mu je ispustiti sve podatke koje ljudi ne mogu primijetiti, smanjiti kvalitetu zvuka i efikasno komprimirati podatke. Advanced Audio Coding (AAC) je drugi popularni format koji se zalaže naprednim kodiranjem zvuka te je stvoren kako bi bio bolja verzija od MP3-a, ali to mu nikad nije pošlo za rukom [17].

Audio formati s kompresijom bez gubitka su stvoreni kao izvorne datoteke koje su komprimirane i manje su od izvornih datoteka, ali kako nemaju gubitaka veće su od tipa s gubitcima [13]. FLAC (Free Lossless Audio Codec format), besplatni format audio kodeka bez gubitka, jedan je od najpopularnijih formata tog tipa i nastao je 2001. godine. Ima mogućnost komprimiranja izvorne datoteke za 60% i još pri tome radi se o besplatnom formatu. I danas

se smatra glavnom alternativom MP3-u za glazbu. ALAC (Apple Lossless Audio Codec), Appleov kodek bez gubitka, je drugi popularni tip koji je razvijen 2004., a besplatan je postao 2011. godine. Nije učinkovit kao FLAC, ali Apple korisnici ga moraju koristiti jer Apple uređaji ne podržavaju FLAC [17].

5. Osnove prepoznavanja govornika

5.1. Prepoznavanja govornika

Tehnike identiteta glasa nisu isto što i prepoznavanje govornika. Postoje mnoge druge tehnike identiteta glasa kao što su glasovni klon i odvajanje glasa, ali prepoznavanje govornika srž je svih tehnika identiteta glasa. S njom je cilj prepoznati tko govori. Sinonimi za prepoznavanje govornika (eng. speaker recognition) su: identifikacija glasa (eng. voice recognition), prepoznavanje glasovnog otiska (eng. voiceprint recognition) i talker recognition [4].

Često se miješa prepoznavanje govornika i govora. Prepoznavanje govornika nam otkriva tko govori, a prepoznavanje govora što se govori [4]. Ono što je omogućilo da se može prepoznavati govornika su glasovi koji se razlikuju od osobe do sobe, u što ulazi veličina i oblik govornih organa, dob, spol, jezik pa čak i životno iskustvo [14].

Tehnika prepoznavanja govornika sadrži nekoliko podjela, a prva je podjela po zadatku. Ona se dijeli na: verifikaciju govornika (eng. Speaker verification-SV) i identifikaciju govornika (speaker identification-SID) [14]. Verifikacija govornika bavi se jednim kandidatom na način da uzima govor nepoznatog govornika s njegovim identitetom i utvrđuje odgovara li navedeni identitet govoru. Većinom se koristi kod aplikacija sustava za buđenje (Wake up systems) ili u sustavima sigurnosti. Način na koji radi je jednostavan, prvo se registrira jedan korisnik sa svojim glasom i identitetom, nakon toga svaki sljedeći korisnik koji proba pristupiti aplikaciji snima svoj glas i ako se glas podudara s registriranim korisnikom dobiva pristup, a ako ne onda ga odbija [4], [14]. Identifikacija govornika bavi se s više korisnika i odgovara na pitanje: „Je li ga jedan od njih izgovorio? Ako da, tko?“. Takav tip se koristi više za aplikacije personalizacije. Vrlo je bitno kod tog tipa da se nikad ne može identificirati svih 7 milijardi ljudi jer bi za to trebala prevelika baza podataka. Uvijek se odredi na početku broj kandidata. Te sustav funkcionira na način da se može više ljudi registrirati i prilikom provjere se provjerava sa svakim od registriranih [4], [8], [14].

Druga podjela je po tekstualnom sadržaju. To je zapravo podjela koja djeluje na podjelu prema zadatku. Pa tako postoje 3 tipa: ovisno o tekstu (text-dependent TD), neovisno o tekstu (text-independent TI) i upitan za tekst (text-prompted TP). Kako je to zapravo podjela koja dijeli dva tipa podjele po zadatku na više tipova dobivamo šest problema: TD-SV, TI-SV, TP-SV, TD-SID, TI-SID i TP-SID [4]. Prepoznavanje govornika ovisno o tekstu još se naziva prepoznavanje govornika s fiksnim tekstom. To pretpostavlja da je izgovorena fraza uvijek ista ili s minimalnim promjenama. To u velikoj mjeri pojednostavljuje problem jer varijacije svakog

fonema su male i treba samo modelirati nekoliko fonema i varijacije duljine zvuka su male pa se može čak pretpostaviti i prozor fiksne duljine. Primjeri takvih slučajeva su riječi buđenja kod Googla „Ok Google“, kod Amazona „Alexa“ i kod Apple-a „Hey Siri“, a koristi se i kao lozinke. Prepoznavanje govornika neovisno o tekstu naziva se još i slobodnom formom prepoznavanja govornika. U ovom slučaju ne postoji pretpostavka što će biti izgovoreno. Prepoznaje se govornik na temelju proizvoljno izgovorenih fraza te je taj problem vrlo koristan, ali i vrlo težak. Razlozi težine su: različiti izrazi između upisa u odnosu na prepoznavanje, razlika između treninga u odnosu na vrijeme izvođenja i nepoznata duljina zvuka. Također postoje dvije kategorizacije: odvisno o jeziku to jest fiksni jezik ili neovisno o jeziku to jest bilo koji jezik. Prepoznavanje govornika s upitom o tekstu je zapravo vrlo sličan model kao ovisno o tekstu. Napravljen je zbog više razine sigurnosti jer postoji mogućnost da netko snimi glas registriranog korisnika te preko snimke uspije otključati podatke. Pa tako kod ovog načina registrirani korisnik snima više svojih izjava te sustav svaki put od korisnika traži neku drugu od snimljenih kako bi potvrdio korisnika [4]. [14].

Trijaža sustava nam govori koji sustav kada odabrati i omogućuje nam smanjenje računalne složenosti. Pa su tako mali i jeftini modeli većinom modeli koji su ovisni o tekstu, dok su veliki i skupi modeli prepoznavanja govornika modeli koji su neovisni o tekstu. Skupi model radimo samo kada je to potrebno. Ako jeftini model ima visoku ocjenu to jest testiranja su pokazala da model uspješno identificira govornika u velikom broju slučajeva, tada ga samouvjereno prihvaćamo i nemamo potrebe za skupim modelom. Ako jeftini model ima nisku ocjenu tada ga samouvjereno odbijamo i nema potrebe za skupim modelom. A ako jeftini model ima srednji rezultat tada postoji nesigurnost i tada se pokreće skupi model za konačnu odluku [4].

5.2. Tijek rada sustava za prepoznavanje govornika

Postoje tri faze rada za sustav prepoznavanja govornika. Prva je faza obuke koja se dešava prije rada modela to jest u procesu razvoja. Nakon nje slijede druge dvije faze to su faza upisa i faza prepoznavanja koje se dešavaju dok je model u izvođenju. Faza obuke ili treninga odvija se tijekom razvoja nakon koje dobivamo model strojnog učenja. Riječ model može biti dvosmislena u sustavu prepoznavanje govornika pa razlikujemo model koji je ovisan o govorniku koji vrši prikaz svakog govornika. I model neovisan o govorniku, a to je globalni model strojnog učenja koji generira ugradnju govornika. Prva faza koja se odvija tokom izvođenja sustava je faza upisa te u njoj svaki govornik daje više audioisječaka za prijavu. Na kraju ove faze korisnici dobivaju svoje profile. I posljednja faza je faza prepoznavanja gdje se vrši provjera i usporedba s unesenim profilima da li je novi govornik unutar baze ili ne [4].

5.3. Procjena i mjerni podatci

Ako imamo dva sustava i želimo znati koji je bolji glavna stvar koja nam to može otkriti su mjerni podatci. Korištenje istog skupa podataka o provjeri protokola čini usporedbu pravednom. Načela koja su bitna tijekom procjene su da nema preklapanja s izgovorima u podacima u obuci i da nema preklapanja s govornicima u podatcima u obuci. Postoje dvije vrste provjera. Procjena temeljena na parovima gdje se određuje popis probnih parova i svaki probni par ima dvije izjave te se stavlja na njih binarna naljepnica kako bi se znalo da su te dvije izjave iz istog govornika. Druga vrsta je vrednovanje temeljeno na skupu koje podrazumijeva da korisnik pruža višestruke izjave o upisu. To vrednovanje sadrži dva podskupa, skup upisa i skup provjere, te oni pokrivaju potpuno iste govornike, ali različite izgovore. Prilikom usporedbi nastale su četiri vrlo važne pogreške za vrednovanje. Istinito prihvaćanje (True accept-TA) nastaje ako je potvrđeno da je isti govornik za dvije izjave i u skupu upisa i u skupu provjera. Lažno odbijanje (False reject-FR) nastaje ako je u skupu upisa potvrđen isti govornik za izjave, a u grupi provjera različit, dok je za lažno prihvaćanje (False accept-FA) suprotno od lažnog odbijanja. I posljednje je Istinito odbijanje (True reject-TR) gdje je različit govornik u oba skupa [4].

Tablica 1 Pogreške za vrednovanje procjena

Skup upisa	Skup provjere	Pogreška
Isti govornik	Isti govornik	Istinito prihvaćanje
Isti govornik	Različit govornik	Lažno odbijanje
Različit govornik	Isti govornik	Lažno prihvaćanje
Različit govornik	Različit govornik	Istinito odbijanje

(Izvor: izrada autora prema Wang, Q. 202*2.)*

Pomoću četiri navedene pogreške nastali su mjerni podatci. Prvi je stopa lažnog prihvaćanja (FAR) ili ti vjerojatnost prihvaćanja lažnog govornika. Računa se na način da se broj lažnog prihvaćanja podjeli s zbrojem broja lažnog prihvaćanja i broja istinitog odbijanja (FAR = FA/(FA + TR)). Druga je stopa lažnog odbacivanja (FRR) ili vjerojatnost odbijanja pravog govornika koja se računa na način da se broj lažnog odbijanja podjeli s zbrojem broja lažnog odbijanja i istinitog prihvaćanja (FRR = FR / (FR + TA)). Kako ne možemo uspoređivati dva sustava temeljem na FAR-u i FRR-u napravljene su dvije krivulje za bolju usporedbu: Krivulja radnih karakteristika prijemnika ROC i krivulja kompromisa s pogreškama pri otkrivanju DET. DET je krivulja koja se više koristi, ona za x os ima FAR, a za y os FRR. Vrijednosti bliže

donjem lijevom kutu predstavljaju bolje performanse. Iz nje je stvorena i jednaka stopa pogrešaka EER (Equal Error Rate) koju crtamo na DET grafu kao dijagonalnu liniju y=x i ona prolazi na mjestu gdje je FAR = FFR i ta vrijednost označuje EER, a što je manji EER to su bolje performanse [4].

5.4. Normalizacija rezultata

Primjenjivanje istog globalnog praga na sve govornike nije moguće iz razloga što neki govornici imaju jedinstvenije glasove, dok drugi imaju češće. Pa normalizacija rezultata ima u cilju smanjiti varijabilnost rezultata. Najpoznatije normalizacije su Z-normalizacija ili nulta i T-normalizacija ili normalizacija testa. Iz nje su stvorene i ostale normalizacije kao što su TZ-normalizacija koja prvo primjenjuje T pa Z-normalizaciju i S-normalizacija koja je simetrična normalizacija. Normalizacija rezultata korak je podešavanja između treninga i procjene [4], [14].

Z-normalizacija sadrži z-rezultat koji nam daje spoznaju koliko je podatkovna točka daleko od srednje vrijednosti. Sama normalizacija nam koristi kako bi mogli napraviti usporedbu s normalnim rezultatima. Prilikom izračuna z-rezultata potrebno je znati srednju vrijednost μ i devijaciju σ. Osnovna formula glasi: $z = (x – \mu) / \sigma$. Gdje z predstavlja z-rezultat, x našu vrijednost, μ srednju vrijednost i σ devijaciju. Pa tako ako imamo našu vrijednost 190, srednju vrijednost 150 i devijaciju 25. z-rezultat bi iznosio 1,6 što bi značilo da je naš rezultat 1,6 standardnih devijacija iznad prosjeka. [20] Često se koristi kod prepoznavanja govornika prilikom rada aplikacije na način da se uzima određeni broj iskaza od različitih govornika i oni se uspoređuju s određenim govornikom i nakon čega se odredi devijacija i srednja vrijednost i na kraju se odredi z-rezultat i na taj način ubrzavamo provjeru.[4].

T-normalizacija se koristi na način da se uzima određeni broj govornika i za svakog od njih radimo bodovanje za svakog govornika posebno i radimo skup kojem odredimo devijaciju i srednju vrijednost. Nakon čega isto bodovanje radimo za određenog govornika. I na kraju Uzimamo govornike koji su najbliži ciljanom govorniku i radimo skup po kojem radimo novu normalizaciju [4].

5.5. Rani pristupi prepoznavanju govornika

Od početaka se koristilo više načina i metoda za identifikaciju govornika. Kroz sljedeće podnaslove proći će se kroz najpopularnije i najbitnije metode i načine koji su korišteni prije samog dubokog učenja.

5.5.1. Gaussova mješavina modela (eng . Gaussian mixture models GMM-a)

Gaussova raspodjela se počela vrlo rano koristit. Jednostavna je za korištenje, vrlo je poznata i sadrži teorem središnje granice CLT koji nam govori da zbroj identičnih slučajnih varijabli teži Gaussovoj raspodjeli čak i ako same izvorne varijable ne slijede Gaussovu distribuciju [15]. Gaussian može predstavljati bilo što: temeljna učestalost neke riječi, format, intenzivnost, omjer signala i buke. Osnovna ideja Gaussiana je da se koristi zbroj Gaussiana za komplicirane raspodjelepa tako Gaussian uvodi matrice kovarijance koje su zapravo kvadratne matrice koje sadrže varijance u glavnoj dijagonali, a kovarijance iznad glavne dijagonale. Pa tako za svakog korisnika svaka komponenta ima svoju matricu kovarijance. Za svakog korisnika sve komponente dijele matricu kovarijanaca i sve komponente korisnika dijele istu matricu. Identifikacija samog govornika slijedi na način da uz dobru procjenu svaki govornik može biti predstavljen pomoću GMM-a, i da se putem njega može pronaći najbližeg govornika među svim kandidatima [4].

5.5.2. Univerzalni pozadinski model (eng. Universal background model UBM)

Kako je GMM otvoreni sustav prepoznavanja govornika moglo je biti puno varalica, a i nije imao mogućnost korištenja zapisnika vjerojatnosti. Također svaki govornik ima svoj GMM pa se tu stvara previše parametara za procjenu i neki govornici nisu imali dovoljno podataka. Kao alternativa je stvoren Universal background model (UBM) kojem su se parametri mogli procijeniti pomoću standardnog algoritma za povećanje očekivanja i maksimiziranja te jednostavno objedinjuje podatke o treningu svakog govornika [14]. Kako je on dio GMM-a česta kratica mu je GMM-UBM iako se koristi i UBM. Primjena se vrši pomoću Bayesove adaptacije gdje smo uzimali GMM i prilagođavali ga UBM-u nakon toga su sve komponente spojene i podržano je brzo bodovanje to jest testiranje korisnika, što daje bolje performanse [4].

5.5.3. Stroj potpornih vektora (eng. Support vector machines SVM)

Stroj potpornih vektora jedan je od najpopularnijih algoritama za klasifikaciju i regresiju i temelji se na načelu maksimalne margine. U prepoznavanju govornika GMM-SVM bio je vrlo uspješan. Postoji linearni i nelinearni pristup. Kako u stvarnim primjenama linearni nije bio

idealan , nelinearni je bio više istaknut pri prepoznavanju govornika jer je koristila pretvorbu s funkcijom jezgre. Stvorene su dvije različite funkcije jezgri koje su funkcionirale na bazi supervektora, te se jedna pokazala boljom od druge, ali u konačnici obje su bile daleko bolje od običnog UBM pristupa [4].

5.5.4. Faktorska analiza i analiza zajedničkog faktora (joint factor analysis)

Dimenzionalnost supervektora može biti previsoka te može sadržavati nepotrebne informacije govornika, kao što su mikrofonske pogreške, akustično okruženje i pozadinska buka, stvorena je faktorska analiza koja je statička metoda koja se fokusirala na opažane metode, dok je neopažene smatrala samo čimbenicima. Ona izvodi procjenu maksimalne vrijednosti matrice opterećenja, može transformirati opažane varijable u Gaussianove latentene varijable i temelji se na razgradnji pojedinačnih vrijednosti.

Analiza zajedničkih faktora je pristup faktorske analize na GMM-UBM. Ona automatski odvaja čimbenike govornika i čimbenike kanala, te samo promatra čimbenike govornika, a čimbenike kanala odbacuje [4].

5.5.5. I-vektor

Idealan slučaj je da čimbenici kanala ne sadrže nikakve informacije vezane uz korisnika to jest govornika, ali eksperimenti su dokazali da čimbenici kanala sadrže jako puno informacija o govorniku te ako ga ne koristimo gubimo velik broj informacija. Pa se iz tog razloga razvio bolji pristup faktorske analize i-vektor, koja je zapravo pojednostavljenje analize zajedničkih faktora. Osnovna ideja je da se ne razlikuju čimbenici govornika i kanala i da to dvoje spajamo kao jedne čimbenike i da su oni poznati kao ukupni čimbenici. Tako da, kada bi usporedili i-vektor s analizom zajedničkog vektora obje se temelje na faktorskoj analizi, samo što je i-vektor učinkovitiji i brži jer ne radi razdvajanje nego odmah sve obrađuje, a pritom daje sigurnije rezultate [4].

6. Osnove dubokog učenja

6.1. Uvod u duboko učenje

Moderni sustavi prepoznavanja govornika temelje se na dubokom učenju. Duboko učenje temelji se na neuronskim mrežama, neuronske mreže su vrsta strojnog učenja. Duboko učenje nije samo važan pristup za prepoznavanje govornika, već se radi i dominantnom pristupu za gotovo sve probleme sa strojnim učenjem, kao što su računalni vid i obrada slike, razumijevanje jezika i prepoznavanje govora. Duboko učenje poznato je po tome što zahtjeva jake računalne komponente iz razloga što se koristi puno podataka s puno raznih usporedbi. Podatci dubokog učenja jednostavni su za proizvodnju i vrlo lako se šire, dok je softver jednostavan za korištenje te čini krivulju učenja glatkom [5].

Neuronske mreže inspirirane su biološkim neuronskim/ živčanim sustavom. Neuron je osnovna funkcionalna jedinica mozga. Poznato je da mozak može obavljati komplicirane zadatke i to više njih u isto vrijeme, ali neuron je relativno jednostavan. On prima signale putem dendrita, zatim provodi električni signal putem aksona i šalje električne signale putem sinapse. Provođenje živčanog impulsa slijedi princip sve ili ništa, pa tako ako neuron reagira, on mora potpuno reagirati, a ne odraditi reakciju do pola. Osnovna jedinica živčane mreže je umjetni neuron [5]. Postoje 4 tipa neurosnkih mrežea: Neuronska mreža feed-forward, konvolucijalna neuronska mreža (CNN), ponavljajuća neuronska mreža (RNN) i neuronske mreže bazirane na pozornosti. Od tipova i neurona možemo izgraditi komplicirane neuronske mreže [4].

6.2. Feed-forward neuronske mreže

Feed-forward neuronska mreža sastoji se od jednostavnih umjetnih neurona i imamo više neurona na istom mjestu. Također se dodaje još jedan dodatni sloj koji uzima izlaz prethodnog sloja. Ovaj tip poznat je i kao višeslojni perceptron (MLP). Funkcija nelinearne aktivacije omogućuje sastavljanje komplicirane funkcije kombiniranjem jednostavnih umjetnih neurona. Opis funkcije je temeljen na biološkim živčanim stanicama sve ili ništa pa tako i sam princip feed-forward slijedi da mora ono što započne odraditi do kraja [5].

6.3. Konvolucijske neuronske mreže (CNN)

Feed-forward neuronske mreže izvrsne su u rukovanju s niskodimenzionalnim značajkama, ali nisu tako dobre s visokodimenzionalnim podatcima kao što su slike s puno piksela, audio s mnogo uzoraka ili videozapis sa slijedom slika. Pa tako da bismo izbjegli eksploziju parametara, možemo bolje iskoristi prostorno ponovljene informacije. Primjer je otkrivanje ljudskog lica na slici, lice može biti bilo gdje na slici, pa se parametri mogu dijeliti na različitim mjestima slike i zato su nam potrebne konvolucijske neuronske mreže. Konvolucijekse mreže sastoje se od konvolucijskih slojeva koji se sastoje od konvolucijske jezgre. Ista jezgra se koristi na različitim mjestima ulaza i proizvodi više izlaza. Konvolucijski sloj ovisi o broju jezgri, veličini jezgri, manja jezgra može se primijeniti na više lokacija i koraku jezgre to jest kolika je udaljenost između dvije jezgre. CNN se često koristi za audio valne oblike, spektograme ili akustične značajke i računalni vid. Konvolucijske neuronske mreže često dolaze s udruživanjem kako bi se smanjile dimenzije značajki i smanjila lokalna osjetljivost [4].

6.4. Rekurentne neuronske mreže (RNN)

U obradi zvuka i govora često se bavimo sekvencijalnim podacima. Sekvencijalni podaci imaju vremensku dimenziju, slijed može biti dug i duljina slijeda može biti nepoznata. Iste informacije mogu se ponoviti na različitim položajima u nizu te se time želi iskoristiti vremenski ponavljane informacije a pritom uštediti na memoriji. Razlika u CNN-u i RNN-u je da CNN ponovno koristi parametre u prostornoj dimenziji to jest istoj jezgri na svim lokacijama, dok RNN ponovno koristi parametre u vremenskoj dimenziji to jest iste parametre za svaki vremenski korak. Najveći problem s RNN-om je da je teško dugo zadržati pamćenje pa je za rješavanje problema predloženo dugo kratkoročno pamćenje (eng. long short-term memory LSTM) te se ono sastoji od ćelija koje prenose informacije o memoriji, ulaznih vrata koji dopušta da unos utječe na memoriju, izlazna vrata koja omogućuju da memorija utječe na izlaz i vrata za zaborav koja baca memoriju. Zatvorena ponavljajuća jedinica (eng. Gated recurrent unit GRU) je drugo rješenje koje je jako slično LSTM-u, ali jednostavnije. Kombinirana su vrata za ulaz i izlaz s jednim ažuriranjem što joj omogućuje da ima manje parametara [4].

6.5. Attention i transformer neuronske mreže

LSTM se bavi problemima RNN-a, ali i dalje ne može riješiti problem dugoročnog pamćenja u potpunosti. Attention i transformer predloženi su izvorno za strojno prevođenje i češće se koristi kao model za prepoznavanje govora nego za prepoznavanje govornika.

Mehanizam Attention-a omogućuje modelu dolazak do bilo koje riječi unutar rečenice ili ti do bilo kojeg dijela u nizu, koristeći mjere relativnosti i pružajući relevantne informacije [4]. Na primjer imamo rečenicu „Ona jede zelenu jabuku.“. Gdje nam je povezanost između jede i zelenu jako mala to jest te dvije riječi nisu povezane, dok s druge strane riječi jede i jabuku imaju veliku povezanost i kao što i jabuka i zelenu imaju veliku povezanost možemo reći da i u konačnici jede i zelenu mogu imati veliku povezanost ovisi na pregled cijelog konteksta rečenice i na taj način funkcionira i attention mehanizam [21].

Prije razvoja transformera modeli su se bazirali na čistom RNN-u, ali dodavanjem mehanizma attentiona performanse modela su se povećale. Transformeri tako predstavljaju model dubokog učenja koji prilagođava mehanizme attention-a i omogućava korištenje istih mehanizama bez RNN-a, a pritom obrađuju sve dijelove niza to jest tokene i izračunavaju težinu attention-a. Transformeri imaju mogućnost izračunati sve tokene paralelno što dodatno ubrzava proces obuke. [21]

Model transformera koristi arhitekturu koder-dekoder. Koder generira kodiranje koje sadrži informacije o tome koji su dijelovi relevantni jedni drugima i prosljeđuje ih sloju kodera kao ulaze. Dekoderi rade suprotno, na način da uzimaju kodiranja i ugrađene informacije koriste za generiranje sekvenci. Svaki sloj kodera i dekodera koristi attention mehanizam. [21]

7. Prepoznavanje govornika pomoću dubokog učenja

I ako se danas sve manje koriste tradicionalni načini pri izgradnji modela za identificiranje govornika i dalje se koriste na neke načine, pa na primjer koriste se zajedno s neuronskim mrežama i tako se dobiju bolji rezultati od korištenja samo tradicionalnih načina. Ipak korištenje izravno samo neuronskih mreža pokazalo je najbolje rezultate i danas se u većini slučajeva koriste modeli s izravnom uporabom neuronskih mreža.

7.1. Neizravna uporaba neuronskih mreža (eng. Indirect use of neural networks)

Neizravna uporaba neuronskih mreža temeljena je na konvencionalnim okvirima: Gaussian mixture models, stroju potpornih vektora i probabilističke linearne diskriminante analize. Izravna uporaba prepoznaje govornika kao problem dubokog učenja i ne oslanja se na druge modele. Tri neizravna pristupa su tandem deep features, DNN i-vektor i j-vektor [4].

Tandem deep features predložen je 2014. godine i koristi neuronske mreže neizravno. Napravljen je na temelju konvencionalnog GMM-UBM okvira i bavi se akustičkim značajkama. Koristi nekoliko principa. Ograničeni Boltzmanov stroj (eng. Restricted Boltzmann machine RBM) koji provjerava je li dubok pristup učenju bez nadzora. Drugi je DNN koji diskriminira telefon. Neuronska mreža koja je obučena za klasificiranje državnih oznaka trifona. Trifon je niz od tri uzastopna fonema. I treći je diskriminirani DNN koji koristi neuronsku mrežu kao klasificiranje oznaka govornika, i po oznakama govornika raspoznaje govornike [4].

DNN i-vektor predložen je 2014. godine i koristi neuronske mreže neizravno te se temelji na konvencionalnom i vektorskom okviru [4]. Metodologija ekstrakcije i-vektora temeljena na faktorskoj analizi mogla bi se smatrati metodom probabilističke kompresije koja smanjuje dimenzionalnost super-vektora u skladu s Gaussovim modelom. Super-vektor govornika i kanala projiciran je u prosotru ukupne varijabilnosti [18].

J-vektor predložen je 2015. godine i također koristi neuronske mreže neizravno. Temelji se na probabilističkoj linearnoj diskriminanti analize (eng. Probabilistic Linear Discriminant Analysis PLDA). Jedna neuronska mreža se koristi za obavljanje dva zadatka: klasificiranje oznaka govornika i klasificiranje fonema. Zbrajaju se unakrsno-entropijski gubici dvaju zadataka i na kraju PLDA klasificira govornika [4].

7.2. Izravna uporaba neuronskih mreža

Neizravna metoda oslanja se na konvencionalne okvire i na mreže za prepoznavanje govora. To je samo zato što duboko učenje u to vrijeme nije bilo zrelo i bio je nedostatak velikih skupova podataka. Korištenje bolje razvijenih podataka i nagomilavanje svih dostupnih značajki tada je bila privremena strategija za osvajanje najboljih performansi. Te je danas duboko učenje sve potrebnije i usredotočeno je na korisnika, a ne na modele prepoznavanja govora [4].

Primjer eksperimenta identifikacije govornika koristeći višeslojnu feed-forward neurosku mrežu u Estoniji napravili su dva znanstvenika Toomas Altosaar i Einar Meister na Institutu za kibernetiku. Oni su napravili sustav od 279 različitih slojeve koji su bili podijeljeni u 3 različite faze: faza koja je primala izjave ili buku, faza koja je odvajala muške i ženske glasove i posljednja je korištenjem neuronske mreže prepoznavanje govornika. [22]

Hossein Salehghaffari je prepoznavanjem govornika uz korištenje konvolucijske neuronske mreže htio prikazati da je to napredniji način od tradicionalnih načina kao što su GMM i i-vektor koji kako on navodi nisu prirodni za prepoznavanje govornika. On je svoj model verifikacije govornika podijelio u tri faze: razvoj, upis i evaluaciju. U fazi razvoja uzimao je izgovore govornika kako bi stvorio pozadinski model pomoću DNN-a. U fazi upisa već postoji model za svaki identitet govornika i cilj je samo po prepoznatom izgovoru sortirati određenog govornika u njegov model putem DNN-a. U posljednjoj fazi to jest fazi evaluacije, testiraju se novi iskazi te se uspoređuju sa svakim do tad unesenim govornikom. Tu se koristi stopa pogreške (EER) i također se utvrđuje stopa lažnog odbijanja i prihvaćanja. I na kraju rezultat se mjeri sličnošću između ispitanog iskaza i određenog govornika. Na kraju je potvrđeno da je CNN pružao bolje performanse od tradicionalnih načina.[23]

Znanstvenici na Sveučilištu u Sheffieldu napravili model prepoznavanja govornika s rekurentnim neuronskim mrežama. Motiviralo ih je to što RNN ima sposobnost obrade kratkoročnih spektralnih značajki u ovoj domeni, ali i reagira na dugoročne vremenske događaje. Oni su u svojoj primjeni uzeli feedforward arhitekturu istu kao i Toomas Altosaar i Einar Meister koji su navedeni prije. Samo što su oni dodali skriveni sloj koji se nalazi između ulaznog i izlaznog stanja. Oni su unutar skrivenog sloja mijenjali broj glavnih i skrivenih jedinica u RNN arhitekturu i ispitivali su učinak na brzinu konvergencije i stope pogreške pri klasifikaciji. U izlaznom sloju koristili su normaliziranu eksponencijalnu funkciju kako bi dobili da zbroj izlaza bude jednak 1. Uvježbavanje je koristilo širenje kroz vrijeme te se za svaki izgovor uspoređivao sa ciljanim izgovorom i radio se trening dok greška izlaza za sve izgovore nije manja od 0. Te

su oni pokazali da RNN se jako dobro ponaša u sustavima s od 8 do 16 govornika. Te su pri radu od 70 različitih grupa od 10 govornika imali prosječnu pogrešku od 3,17%. [24]

7.3. Funkcija gubitka

Postoje dvije vrste funkcije gubitka: prepoznavanje govornika kao problem klasifikacije više govornika to jest modeliranje identifikacije mnogih govornika i prepoznavanje govornika kao binarni problem to jest modeliranje provjere jednog govornika [4].

Klasifikacija je jedan od najtipičnijih problema strojnog učenja, na primjer kod slike se pita koji je to objekt, kod prepoznavanja govornika pitamo tko je osoba koja govori. Pretpostavljamo da je svaki govornik zasebna kategorija. Tako obukom svaki govornik ima globalno jedinstvenu oznaku klase. Svaka izreka ima oznaku korisnika. Ako dvije izjave imaju istu oznaku korisnika, one pripadaju istom korisniku. Tijekom rada govornik se može razlikovati od bilo koga u podacima u obuci [8]. Križna entropija gubitka ili unakrsna entropija gubitka temelji se na teoriji informacija. Sloj softmax prima podatke od značajki neuronske mreže, a unakrsna entropija se koristi kao funkcija gubitka. "Softmax gubitak" je uobičajeni naziv za ovaj gubitak. Ali za zadatke klasifikacije (kao što je postavljanje uzoraka u zadanu klasu), softmax gubitak je veći i prikladniji. Verifikacija je, za razliku od kategorizacije, otvoren zadatak. [19]. Prilikom rada koriste se ugradnje za provjeru gubitaka. Problem je što križna entropija čini trening neučinkovitim jer dolazi do velikog gubljenja podataka i trening se jako razlikuje s vremenom izvođenja [4].

S obzirom na dva ulaza stvaramo binarno predviđanje a to su jesu li njihove etikete različite ili iste. Potrebno je napraviti po nekoliko ulaza za svaki korak treninga. Binarna križna entropija ima problem s klasifikacijom, ali daje dobro rješenje. Pa je stvoren trostruki gubitak [4]. Za trening koji se temelji na gubitku tripleta neophodan je temeljit, dugotrajan proces odabira tripleta koji je osjetljiv na performanse. Generiranje tripleta iz mini-serija i korištenje softmax predvježbanja dvije su vrijedne pažnje inicijative za poboljšanje treninga temeljenog na gubitku tripleta [19]. On daje ili pozitivno ili negativno ispitivanje. Izazov predstavlja uravnotežiti pozitivna i negativna ispitivanja tijekom treninga. Pa se umjesto para izjava za provjeru koriste tri izjave. Jedna izjava se naziva sidro i s njom se uspoređuje drugo je pozitivna izreka koja je od istog korisnika, a treća je negativna izreka koja je od drugog korisnika. Proces nalaže da se pozitivna izreka povuče prema sidru, a negativna odgurne. Ovaj sustav se većinom koristi za prepoznavanje lica. Osnovno načelo je da se izabere najteža trojka. A izazov predstavlja što se parametri mreže mijenjaju tijekom treninga, ne može se na početku odabrati optimalne trojke, pa se svakih nekoliko koraka mora ponovno odabirati optimalne trojke i svaki put kad se biraju nove trojke taj proces se zove mini serija [4].

End-to-end sustav nije standardna definicija cjelovitog sustava dubokog učenja. Prvi end obično se odnosi na ulaz sustava to jest zvuk, a drugi end na izlaz sustava to jest predviđanje. Obično end-to-end sustavi dubokog učenja ne koriste druge modele osim neuronskih mreža. Ima samo jedan model neuronske mreže i optimizira samo funkciju jednog gubitka [4]. Pristup end-to-end izravno trenira diskriminativni model koji je najčešće neuronska mreža, kako bi što bolje usavršila diskriminativno ugrađivanje govornika dok se pri tome koriste različite strukture neuronskih mreža. Često se koriste ponavljajuće neuronske mreže (kratica RNN) za izdvajanje značajki identiteta na način da se uvode razni mehanizmi pažnje putem kojih se poboljšava sustav provjere identiteta govornika [19]. Postoji više faza end-to-end sustave. U fazi upisa imaju više upisnih izjava i u fazi prepoznavanja može postojati ili ciljani govornik ili varalica. Tijekom treninga koristi torbu s N+1 izgovora. N broj izjava predstavlja broj izjava za upis govornika, a jedna dodatna je za procjenu govornika ili varalice. Procjena može samo biti ili pozitivna ili negativna. Kako bi izabrali dobar broj izjava treba simulirati logiku vremena izvođenja, pa tako broj izjava ovisi o vremenu izvođenja aplikacije. Kada izračunamo prosjek koristimo slučajni podskup izgovora [4]. End-to-end sistemi imaju isti problem kao i gubitci u paru a to je da je obično više negativnih uzoraka nego pozitivnih. Pa je stvorena generalizacija end-to-end gubitka koje koristi načelo maksimalne margine slično stroju potpornih vektora (SVM) [9]. Ideja je da obuka bude temeljena na skupu i da se negativna ispitivanja gledaju samo ona koja su bila blizu da budu pozitivna. Način je se pokazao vrlo učinkovitim [4].

8. Zaključak

Ovim radom prikazano je kako se ljudi mogu identificirati ne samo otiskom prsta ili crtama lica nego i na druge biometrijske načine, a u ovom radu konkretno pomoću glasa. Kako bi se mogli baviti identifikacijom govornika pomoću glasa potrebno je razumjeti kako sam glas funkcionira i što zapravo je glas, te na koji način izvesti promjenu iz digitalnog u analogni oblik. Dok duboko učenje još nije bilo dovoljno razvijeno ili još nije postojalo tradicionalni načini implementacije identifikacije govornika bili su jedini način izrade modela i za to vrijeme su bili konkretni i davali su prilično dobre rezultate sa sigurnim provjerama. Pa se i dan danas mogu naći primjeri gdje se još koriste tradicionalni načini i ako se danas više koriste uz neuronske mreže kako bi dali bolje rezultate od običnih tradicionalnih implementacija. Ali kako je znanost napredovala napredovalo je i duboko učenje i danas modeli identifikacije govornika putem dubokog učenja predstavljaju najbolji način implementacije jer daju najbolje rezultate, najbrže odrađuju zadane zadatke i imaju najvišu stopu sigurnosti rezultata to jest najnižu stopu pogreške.

Danas je daleko poznatije korištenje drugih identifikacija ljudi kao što su već navedena identifikacija putem otiska prsta, ali sve više se koristi identifikacija govornika te će vrlo vjerojatno kroz budućnost i razvijanje tehnologije još više doći do isticanja. Jer kako je otisak prsta za svaku osobu jedinstven tako je i glas svake osobe jedinstven i može se koristit na jednak način za identifikaciju i sigurnost ljudi.

Literatura

[1] IDR&D (2022). What is Voice Biometrics and why should you use it? Dostupno na:

https://www.idrnd.ai/voice-biometrics/ [pristupljeno 10. srpnja 2022.]

[2] Omni intelligence (2019). Voice Biomterics. Dostupno na:

https://omniintelligence.online/voice-biometrics/?gclid=Cj0KCQjwlK-WBhDjARIsAO2sErSf6xO67NibN8K_8NVaNeFQNnknCw-x_yMwdrDjR3zb9B7TNb4DD5MaAl8sEALw_wcB [pristupljeno 10. srpnja 2022.]

[3] Radmilović, Ž. (2008). Biometrijska identifikacija. Dostupno na:

https://hrcak.srce.hr/file/117825 [pristupljeno 10. srpnja 2022.]

[4] Wang, Q. (2022). Speaker Recognition. Dostupno na:

https://www.udemy.com/course/speaker-recognition/ [pristupljeno 10. srpnja 2022.]

[5] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press. Dostupno na:

https://www.deeplearningbook.org/ [pristupljeno 10. kolovoza 2022.]

[6] Biometrija. URL: https://www.cis.hr/www.edicija/LinkedDocuments/CCERT-PUBDOC-2006-11-167.pdf

[7] Kovač, Zlatko. Prepoznavanje osobe u forenzici na temelju glasa. // Policija i

sigurnost 21, 2 (2012). URL: https://hrcak.srce.hr/87241

[8] Jia Y., Yhang Y., Weiss J. R., Wang Q., Shen J., Ren F., Chen Z., Nguyen P., Pang R. Moreno I. L., Wu Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. Dostupno na: https://proceedings.neurips.cc/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf [pristupljeno 10. kolovoza 2022.]

[9] Wat L., Wang Q., Papir A., Moreno I. L (2018). GENERALIZED END-TO-END LOSS FOR SPEAKER VERIFICATION. Dostupno na: https://arxiv.org/pdf/1710.10467.pdf [pristupljeno 10. kolovoza 2022.]

[10] zvuk. Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža, (2021). Dostupno na: https://www.enciklopedija.hr/natuknica.aspx?ID=67594 [Pristupljeno 10. kolovoza 2022.]

[11] govor. *Hrvatska enciklopedija, mrežno izdanje.* Leksikografski zavod Miroslav Krleža, (2021). Dostupno na: https://www.enciklopedija.hr/natuknica.aspx?ID=22886 [Pristupljeno 10. kolovoza 2022.]

[12] Edutorij.e-skole.hr (2020). Zvuk. Dostupno na: https://edutorij.e-skole.hr/share/proxy/alfresco-noauth/edutorij/api/proxy-guest/6b9de2eb-c6d7-412b-8afc-c0820325b64d/zvuk-1.html [Pristupljeno 15. kolovoza 2022.]

[13] Zvuk u multimediji, Audio formati i kvaliteta zapisa, Osnovna škola Antuna Bauera Vukovar, Dostupno na: http://os-abauera-vu.skole.hr/upload/os-abauera-vu/images/static3/1950/File/Audio%20formati%20i%20kvaliteta%20zapisa.pdf [Pristupljeno 15. kolovoza 2022.]

[14] Furui S., (2008) Speaker recognition, Scholarpedia. Dostupno na: http://www.scholarpedia.org/article/Speaker_recognition [Pristupljeno 10. kolovoza 2022.]

[15] Fabien M., Speaker Verification using Gaussian Mixture Model (GMM-UBM), Dostupno na: https://maelfabien.github.io/machinelearning/Speech1/# [Pristupljeno 12. kolovoza 2022.]

[16] Analogno-digitalna pretvorba, TINF, Dostupno na: http://lab425.fesb.hr/TINF/teorijainf_11.htm [Pristupljeno: 29. Kolovoza 2022.]

[17] Lucas M., 10 najčešćih audio formata koji trebate koristiti?, UneDose Dostupno na: https://hr.unedose.fr/article/the-10-most-common-audio-formats-which-one-should-you-use [Pristupljeno: 29. Kolovoza 2022.]

[18] Reyes-Diaz F. J., Hernandez-Sierra G., Calvo de Lara J.R., DNN and i-vector combined method for speaker recognition on multi-variability environments, Dostupno na: https://link.springer.com/article/10.1007/s10772-021-09796-1#author-information [Pristupljeno: 29. Kolovoza 2022.]

[19] Li Y., Gao F., Ou Z., Sun J., Angular Softmax Loss for End-to-end Speaker Verification, Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, Beijing, China, Dostupno na: https://arxiv.org/pdf/1806.03464.pdf [Pristupljeno: 30. Kolovoza 2022.]

[20] Glen S., Z-Score: Definition, Formula and Calculation, StatisticsHowTo.com, Dostupno na: https://www.statisticshowto.com/probability-and-statistics/z-score/ [Pristupljeno: 7. Rujna 2022.]

[21] Alammar J., The Illustrated Transformer, (2018.), Dostupno na: http://jalammar.github.io/illustrated-transformer/ [Pristupljeno: 8. Rujna 2022.]

[22] Altosaar T., Meister E., Speaker recognition experiments in Estonian using multi-layer feed-forward neutal nets, Helsinki University of Technology (1995), Dostupno na: https://www.isca-speech.org/archive/pdfs/eurospeech_1995/altosaar95_eurospeech.pdf [Pristupljeno: 8. Rujna 2022.]

[23] Salehghaffari H., Speaker Verification using Convolutional Neural Networks, NYU Tandon School of Engineering (Polytechnic Institute), New York (2018.), Dostupno na: https://arxiv.org/pdf/1803.05427.pdf [Pristupljeno: 8. Rujna 2022.]

[24] Parveen S., Qadeer A., Green P., Speaker Recognition with Recurrent neural networks, University of Sheffield, Sheffield, UK (2000), Dostupno na: https://www.researchgate.net/publication/221478141_Speaker_recognition_with_recurrent_neural_networks [Pristupljeno: 8. Rujna 2022.]

Popis slika

Popis tablica