Prepoznavanje emocija iz snimke glasa

Langus, Robert

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Robert Langus

PREPOZNAVANJE EMOCIJA IZ SNIMKE GLASA

DIPLOMSKI RAD

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Robert Langus

Matični broj: 44844/16–R

Studij: Informacijsko i programsko inženjerstvo

PREPOZNAVANJE EMOCIJA IZ SNIMKE GLASA

DIPLOMSKI RAD

Mentorica:

Izv.prof.dr.sc. Petra Grd

Varaždin, veljača 2024.

Robert Langus

Izjava o izvornosti

Izjavljujem da je moj 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 potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad bavi se istraživanjem i analizom sposobnosti tehnologije identificiranja i klasificiranja ljudskih emocija na temelju biometrijskih podataka dobivenih iz glasovnih snimaka. U uvodnom dijelu rada, daje se pregled osnovnih pojmova biometrije općenito, kao i biometrije glasa te se istražuje važnost i primjena prepoznavanja emocija u različitim područjima, poput sigurnosti i razvoja sigurnosnih značajka na pametnim uređajima. Rad detaljno opisuje i uspoređuje različite teorijske i metodološke pristupe prepoznavanju emocija, uključujući strojno učenje, duboko učenje i analizu glasovnih karakteristika, kao što su ton, intenzitet i brzina govora. U praktičnom dijelu rada predstavljena je implementacija modela za prepoznavanje emocija iz snimke glasa s pomoću python programskog jezika, temeljena na analizi i obradi glasovnih snimaka.

Ključne riječi: prepoznavanje emocija; biometrija glasa; strojno učenje; python; klasifikacija emocija; snimka glasa

Sadržaj

1. Uvod

U današnjem suvremenom svijetu, proučavanje biometrije predstavlja revolucionarni korak prema razumijevanju kako tehnološke inovacije mogu služiti kao most između identifikacijskih karakteristika pojedinaca i digitalnog svijeta. Biometrija, definirana kroz svoju sposobnost da koristi fizičke ili ponašajne atribute za prepoznavanje i autentifikaciju osoba, našla je svoje mjesto za znanstvena istraživanja i tehnološki napredak. Upravo o tome će biti govor u ovom diplomskom radu. U njemu će biti naglasak na istraživanja i razvoj tehnologija koje omogućuju identifikaciju i interpretaciju ljudskih emocija temeljenih na analizi glasovnih snimaka. Ova sofisticirana područja istraživanja pružaju osnovni temelj za razumijevanje kompleksnih interakcija između govora i emocija, otvarajući put za razvoj naprednih aplikacija sigurnosnih sustava, razne primjene u kliničkoj psihologiji, korisničkoj podršci te neizostavnoj umjetnoj inteligenciji, o kojoj se priča svakodnevno. S obzirom na rastući interes za umjetnu inteligenciju kao i emocionalnu inteligenciju u tehnološkim sustavima, ovaj rad predstavlja analizu biometrije glasa kao ključnog alata u prepoznavanju emocija iz snimke glasa, istražujući njezine metodološke, teorijske i praktične dijelove.

Kroz detaljan pregled literature i postojećih istraživanja u ovom radu bit će približena tema biometrije, kao i koncept biometrije glasa, objašnjavajući osnovne principe i primjene tehnologije prepoznavanja emocija iz snimke glasa. Za bolje razumijevanje funkcionalnosti prepoznavanja emocija bitno je i znanje fiziologije govornog sustava i načina na koji glas prenosi emocije. Osim toga bit će govora o karakteristikama glasa koje su ključne za biometrijsku identifikaciju - uključujući ton, intonaciju, ritam i brzinu govora. Sve navedeno tvori jedinstveni otisak emocionalnog stanja osobe. Rad će pokriti pristupe i tehnike prepoznavanja emocija iz glasa, uključujući tradicionalne metode i one koje koriste napredne tehnike strojnog i dubokog učenja. Nadalje, bit će osvrt na primjene spomenutih tehnologija u stvarnom svijetu, od sigurnosnih sustava do raznih aplikacija te njihov potencijal i izazove s kojima se suočavaju istraživači, programeri i svi oni koji imaju doticaj s osobama gdje je analiza glasa ključan dio poslovanja. U radu će biti i opisana detaljna implementacija jednostavnog sustava za prepoznavanje emocija iz snimke glasa koristeći se python programskim okruženjem. Sve to s ciljem isticanja kako napredak u prepoznavanju emocija iz snimke glasa pruža mogućnosti za poboljšanje interakcije između čovjeka i računala, kao i za unapređenje kvalitete života kroz različite primjene.

2. Metode i tehnike rada

Velika većina teorijskog dijela ovog rada zasniva se na prikupljanju i analizi znanstvenih članaka na temu biometrije i prepoznavanja emocija iz snimke glasa. S obzirom na to da postoji više metoda za prepoznavanje emocija iz snimke glasa napravljena je i usporedba tehnika. Za implementaciju sustava za prepoznavanje emocija iz snimke glasa korišteno je python programsko okruženje. Od biblioteka korištene su librosa, numpy, os, pickle, tensorflow, matplotlib, seaborn te sklearn. Dok od tehnika za prepoznavanje emocija iz snimke glasa korišteni su algoritmi k-najbližih susjeda te konvolucijske neuronske mreže.

3. Uvod u biometriju

Za bolje razumijevanje same teme, prije svega, potrebno je predstaviti pojam biometrije. Biometrija je znanstvena disciplina koja je usmjerena na uspostavljanje identiteta pojedinca na temelju njegovih fizičkih, kemijskih ili ponašajnih karakteristika [1]. Zbog toga pronalazi sve veći interes i potrebu za sustavima upravljanja identitetima velikih razmjera, a upravo funkcionalnost tih sustava počiva na preciznom utvrđivanju identiteta osoba. Zato u današnjem svijetu možemo vidjeti kako ti sustavi postaju sve pristupačniji i izraženiji. Od CCTV kamera do raznih autentifikacijskih sustava, kao što je i čitanje otiska prsta na mobilnim uređajima. Biometrijski sustavi nalaze primjenu u raznim područjima, od informatike i informacijskih sustava, dijeljenja ili zabrane pristupa različitih resursa pa do izvođenja financijskih transakcija na daljinu [2]. Širenjem i prihvaćanjem digitalizacije, uvođenjem raznih sustava na transakcijskoj razini i specificiranih decentraliziranih centara za prikupljanje podataka, dodatno naglašava potrebu za pouzdanim sustavima upravljanja identitetom koji mogu analizirati i upravljati velikim brojem pojedinaca.

Primarni zadatak sustava za upravljanje identitetom je određivanje identiteta osobe ili pak tvrdnje da je osoba uistinu ta za koju se predstavlja. Različiti su razlozi zašto bi ovakva akcija bila potrebna. Međutim fokus je, bar u većini aplikacija, na sprječavanju pristupa zaštićenim resursima od osoba koje nemaju prava pristupa tim resursima. Tradicionalne metode uključuju neku vrstu autentifikacije koristeći se lozinkom ili predmetima kao što su to RFID kartice ili tagovi. Iako su takvi mehanizmi sigurni, vrlo lako se može manipulirati njima. Primjerice lozinka se lako može podijeliti s drugim ljudima, RFID kartice mogu biti ukradene ili se osoba može nalaziti u situacije gdje je primorana dati pristupne podatke zbog ucjene ili sličnih situacija. S obzirom na to da se takvim mehanizmima može lako manipulirati, takvi sustavi postaju kompromitirajući, te se gubi namjeravana sigurnost. Dok s druge strane biometrija nudi prirodno i pouzdano rješenje za određene aspekte upravljanja identitetom koristeći potpuno automatizirane ili polu automatizirane sheme za prepoznavanje pojedinaca na temelju njihovih bioloških karakteristika. Na temelju toga moguće je potvrditi identitet onako kako jest, a ne na temelju onoga što osoba posjeduje. Nije nužno da se biometrija koristi zasebno, ona može biti nadopuna spomenutim mehanizmima, te dodati dodatni sloj zaštite aplikacije. Primjer takve autentifikacije je autentifikacija u dva koraka, koja se danas iz već navedenih razloga sve više preporučuje unutar raznih aplikacija. Učinkovitost takvih autentifikatora bilo biometrijskih ili ne biometrijskih temelji se na relevantnosti za određeni slučaj korištenja, kao i otpornost na različite vrste zlonamjernih napada. Biometrija nudi određene prednosti, poput negativnog prepoznavanja te ne poricanja, koje tokeni i lozinke ne

mogu pružiti. Biometrijski sustavi koriste različite fizičke ili ponašajne karakteristike, uključujući otisak prsta, lice, geometriju ruke/prsta, šarenicu oka, mrežnicu, potpis, način hodanja, dlan, uzorak glasa, uho, vene na ruci, miris ili DNA informacije pojedinca za uspostavljanje identiteta. Iako biometrijski sustavi imaju svoja ograničenja, oni imaju prednost nad tradicionalnim metodama sigurnosti jer se ne mogu lako ukrasti ili podijeliti. Osim što pojačavaju sigurnost, biometrijski sustavi također povećavaju praktičnost za korisnike, smanjujući potrebu za osmišljavanjem i pamćenjem lozinki.[1]

3.1. Karakteristike biometrije i njezina važnost

Prilikom odabira biometrijskog sustava za identifikaciju osoba prvotno pitanje koje se nameće jest – „koja je biometrija najbolja?“. Kao i u većini slučajeva u životu odgovor nije tako jednostavan i direktan. Prema tome, 'najbolja' biometrija ne postoji, već je pitanje što se želi postići, za što će se sustav koristiti i pod kojim uvjetima. Kako navodi A. K. Jain [1] metodologija koja dobro funkcionira u kontroliranom i stalnom uredskom okruženju, kao što je čitač otiska prsta, možda neće biti prikladna za prometnu javnu zračnu luku kroz koji prolaze na tisuće puta više ljudi u manjem razdoblju, ili primjerice proizvodni pogon gdje zbog zaštite na radu, zaposlenici moraju nositi zaštitni materijal kao što su zaštitne rukavice. Potrebno je pažljivo razmotriti primjenu, razumjeti situaciju iz perspektive krajnjeg korisnika te biti svjesni o prednostima uvođenja odabrane biometrijske tehnologije u određeni poslovni proces. Naravno tu nije kraj, nakon uvođenja nove tehnologije, moguće je da se pokaže kako odabrana biometrija ipak ne zadovoljava uvjete te je zato potrebno provesti ispitivanja te evaluaciju. Dodatno je moguće promijeniti dio hardware-a ili software-a, napraviti ponovnu evaluaciju i zaključiti pridonosi li odabrana biometrija boljem poslovnom procesu. [2]

Dakle kada govorimo o biometriji, moramo znati i koje su zadovoljavajuće karakteristike prilikom uporabe neke od tehnologija. Dosad smo vidjeli kako postoje različite primjene biometrije, o kojima će biti detaljnije rečeno u sljedećem poglavlju, te svaka od njih ima svoje prednosti i mane, stoga izbor biometrijske karakteristike za određenu primjenu ovisi o raznim faktorima osim o njenoj sposobnosti podudaranja. A. K. Jain i ostali [3], su identificirali sedam faktora koji određuju prikladnost fizičke ili ponašajne karakteristike za upotrebu u biometriji:

1. Univerzalnost: Svaki pojedinac koji pristupa aplikaciji trebao bi posjedovati tu karakteristiku.
2. Jedinstvenost: Dana karakteristika trebala bi biti dovoljno različita među pojedincima unutar populacije.

3. Stalnost: Biometrijska karakteristika pojedinca trebala bi biti dovoljno nepromjenjiva tijekom vremena u odnosu na algoritam za podudaranje. Karakteristika koja se značajno mijenja s vremenom nije korisna biometrija.
4. Mjerljivost: Biometrijska karakteristika trebala bi se moći prikupiti i digitalizirati s pomoću odgovarajućih uređaja koji ne uzrokuju pretjerane neugodnosti pojedincu. Nadalje, prikupljeni podaci trebali bi biti pogodni za obradu kako bi se iz njih izvukli reprezentativni skupovi značajki.
5. Učinkovitost: Točnost prepoznavanja i resursi potrebni za postizanje te točnosti trebali bi zadovoljiti ograničenja nametnuta primjenom.
6. Prihvatljivost: Pojedinci iz ciljane populacije koji će koristiti aplikaciju trebali bi biti voljni predstaviti svoju biometrijsku karakteristiku sustavu.
7. Zaobilaženje: To se odnosi na lakoću s kojom se karakteristika pojedinca može imitirati koristeći artefakte (npr., lažne prste), u slučaju fizičkih karakteristika, i mimikriju, u slučaju ponašajnih karakteristika.

Biometrijski identifikator	Univerzalnost	Jedinstvenost	Trajnost	Mjerljivost	Učinkovitost	Prihvatljivost	Otpornost
DNA	V	V	V	N	V	N	N
Uho	S	S	V	S	S	V	S
Lice	V	N	S	V	N	V	S
Facialni termogram	V	V	N	V	S	V	N
Otisak prsta	S	V	V	S	V	S	S
Hod	S	N	N	V	N	V	S
Geometrija šake	S	S	S	V	S	S	S
Vena šake	S	S	S	S	S	S	N
Iris	V	V	V	S	V	N	N
Način tipkanja	N	N	N	S	N	S	S
Miris	V	V	V	N	N	S	V
Otisak dlana	S	V	V	S	V	S	S
Retina	V	V	S	N	V	N	N
Potpis	N	N	N	V	N	V	V
Glas	S	N	N	S	N	V	V

Slika 1. Komparacija raznih biometrija prema karakteristikama V - visoko, S – srednje, N - nisko, prema autoru. [4]

Niti od jedne biometrije se ne očekuje da će zadovoljiti sve karakteristike i zahtjeve poslovnog proces, tj. upotpuniti poslovnu logiku. Što potvrđuje odgovor na prvotno pitanje, a to jest da nema idealne biometrije, međutim naći će se ona koja će biti prihvatljiva za dane parametre. Relevantnost određene biometrije za aplikaciju utvrđuje se ovisno o prirodi i zahtjevima aplikacije te svojstvima biometrijske karakteristike.

3.2. Primjene biometrije

Određivanje identiteta osobe s visokom preciznošću, u međusobno povezanom društvu, postaje ključan element u velikom broju primjena. Sve se češće postavljaju pitanja kao "je li ta osoba uistinu ona za koju se izdaje?", "je li ova osoba ovlaštena biti u ovom prostoru?", te mnoga druga ovisno o scenariju. Zbog sve većih sigurnosnih zabrinutosti i brzog napretka u području mrežnih tehnologija, komunikacije i mobilnosti, potreba za pouzdanim tehnikama autentifikacije korisnika naglo je porasla. Stoga se biometrija sve više koristi u različitim primjenama, koje se mogu podijeliti u tri glavne kategorije, navodi A. K. Jain [1]:

1. Upotreba u komercijalne svrhe - uključuje mrežne prijave, sigurnost elektroničkih podataka, pristup internetu, bankomati, fizička kontrola pristupa, mobilni uređaji, medicinski zapisi

2. Upotreba u vladine svrhe - osobne iskaznice, vozačke dozvole, putovnice, kontrola granica

3. Upotreba u svrhe forenzike - identificiranje tijela, kriminalne istrage, utvrđivanje očinstva/majčinstva

U nastavku slijedi i nekoliko detaljnijih primjena biometrije za navedene kategorije. Što se tiče komercijalne svrhe možemo vidjeti da danas gotovo svaki pametni telefon ima mogućnost otključavanja zaslona otiskom prsta ili prepoznavanjem lica. Nadalje, većina bankomata ima ugrađene kamere. Osim za snimanje, moguće je i identificirati slučajeve ponašanja, odnosno bihevioralne karakteristike biometrije kod korisnika, koje odstupaju od standarda. Primjerice sustav može označiti transakciju kao sumnjivu, ako primijeti da korisnik nelagodno podiže gotovinu ili često gleda preko ramena. U poslovne svrhe čitač otiska prsta može biti kao identifikator da je osoba u određeno vrijeme na određeni dan bila na poslu, te se na kraju radnog vremena odjavila. Za razliku od komercijalnih primjena, primjene u vladine svrhe su na višem nivou te pouzdanije. Najpoznatiji primjer kod nas su vjerojatno biometrijske putovnice, koje osim promijenjenog izgleda sadrže digitaliziranu sliku lica, otiske dvaju kažiprsta te čip s osobnim podacima osobe. Iako kod nas nije uobičajena digitalizirana provjera identiteta već provjera od strane ovlaštenog osoblja na način pregleda lica i putovnice te utvrđivanja radi li se o istoj osobi sa slike. Zbog većeg kapaciteta i protoka u većim milijunskim gradovima kao što je London, prilikom ulaska u državu skenira se putovnica te se gleda u kameru. Sustav automatikom provjerava odgovara li slika osobe onoj u putovnici. Na taj način ubrzava se protočnost i smanjuje se ljudska pogreška. Što se tiče biometrije u forenzici,

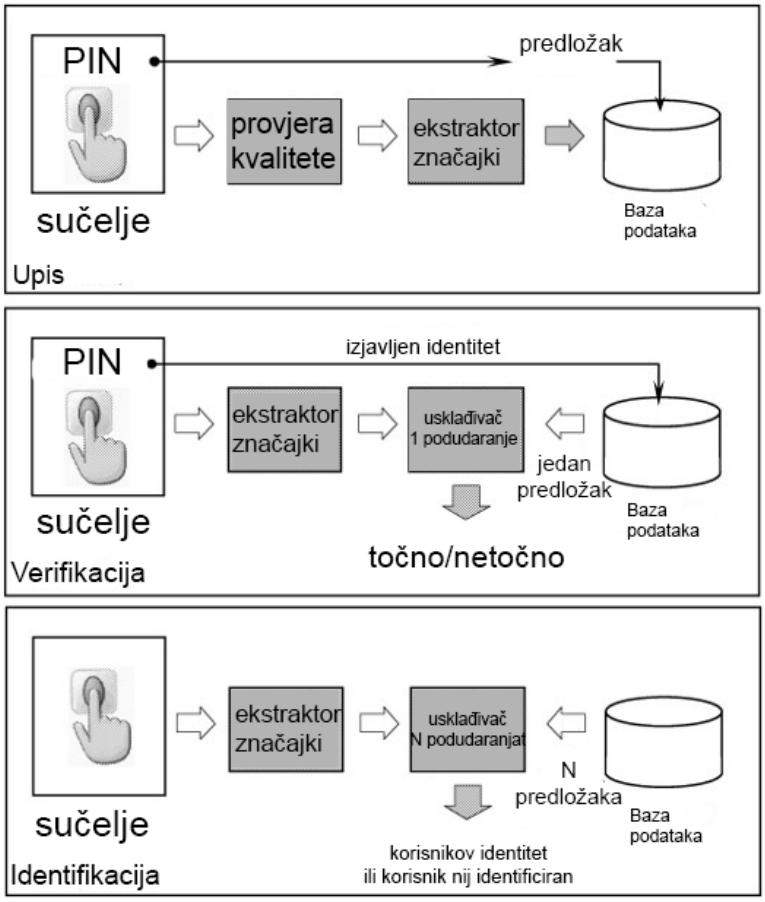
vjerojatno je opće poznato iz koje kakvih kriminalističkih serija, da se osobu može identificirati s pomoću bilo kojeg dijela ako uzmemo u obzir njen DNA. Naravno, osim spomenutih triju glavnih kategorija postoje još deseci, ako ne i stotine, drugih potencijalnih primjena koje nisu obuhvaćene u ovom radu, a neke od njih specifične su za organizacije i autoritete koji žele poboljšati pouzdanost verifikacije identiteta ili jednostavno automatizirati proces. Moguće su i snažne marketinške inicijative gdje proizvođači ili pružatelji usluga mogu pružiti dodatne koristi korisnicima putem usvajanja biometrijske tehnologije. Što ćemo još u budućnosti vidjeti ostaje samo na kreativnosti inovatora.

3.3. Razlika između verifikacije i identifikacije

Kad govorimo o biometriji bitno je spomenuti i dvije vrste operativnosti samog sustava – to su verifikacija i identifikacija. Biometrijski sustavi mogu koristiti biometrijske predloške, koji utječu na dizajn sustava kao i pitanja uspostavljanja kriterija učinkovitosti istog [1]. Bitno je poznavanje terminologije i tematike, ako razmatramo upotrebu biometrije. Osim toga, ovi čimbenici trebali bi se uzeti u obzir prije detaljnog planiranja dizajna biometrijskih sustava. Pojmovi identifikacije i verifikacije često se pogrešno koriste prilikom razmatranja biometrijskih sustava [2]. Trenutno većina biometrijskih sustava radi u modu verifikacije. S obzirom na to da se navedene dvije operacije značajno razlikuju, bitno je shvatiti način funkcioniranja pojedine operativnosti.

Verifikacijski biometrijski sustav koriste predloške koji su povezani s korisnicima. Uobičajeno povezivanje je putem referenciranja na korisnički broj, odnosno zapis u bazi podataka biometrijskog sustava. Prilikom transakcije, osoba prvo dohvaća predložak iz baze te nakon toga prislanja prst na čitač otiska prsta, točnije daje neki od svojih biometrijskih uzoraka [1]. Zatim sustav provodi usporedbu čiji je rezultat binarno stanje – true ili false. Razmatrajući navedene činjenice, zadatak takvog biometrijskog sustava nije pretjerano težak. Sustav uspoređuje dva skupa informacija prema unaprijed definiranim pragom podudarnosti. Verifikacija se tipično koristi u aplikacijama gdje je cilj spriječiti neautorizirane osobe u korištenju usluga. Formalno gledano, verifikaciju možemo postaviti kao problem klasifikacije u dvije kategorije: dati tvrđeni identitet (I) i skup značajki upita (xA), potrebno je odlučiti pripada li (I,xA) klasi "autentičnih" ili "prevaranata". Odluka se temelji na predefiniranom pragu η, a sustav dopušta pristup uslugama ako je tvrdnja identiteta ocijenjena kao "autentična" [3]. Za razliku od operativnosti verifikacije, identifikacijski biometrijski sustav radi na principu da osoba daje svoj biometrijski uzorak, te očekuje od sustava da je prepozna kao ne/ovlaštenog korisnika. Ovdje možemo vidjeti kako takav sustav mora proći kroz cijelu svoju bazu podataka

predložaka, usporediti je s danim uzorkom te vratiti specifičnu korisničku referencu, što je znatno kompleksniji proces. Već je tu vidljivo da može doći do problema ako je baza predložaka poprilično velika, a time i podudarnosti tipa „false positive“ mogu postati veće, poput FBI-jevog integriranog automatiziranog sustava za identifikaciju otisaka prstiju (IAFIS), koji ima otprilike 60 milijuna upisanih korisnika. Zato je bitno utemeljiti dobar algoritam za napredno procesiranje i čim bolje rasuđivanje biometrijskih uzoraka. [2]



Slika 2. Prikaz dijagrama procesa verifikacije i identifikacije. [4]

Od velike važnosti je i dobro upravljanje biometrijskim predlošcima. Kvaliteta i jedinstvenost predloška izravno utječu na sposobnost sustava da ispravno identificira ili verificira osobu. Kao i kod izbora biometrije ne postoji jednoglasno definirano koji bi sustav bio bolji, nego ponovo ovisi od situacije i što se sa sustavom želi postići. Međutim postoji praktično pravilo koje objašnjava prednosti i nedostatke: Sustavi poput skeniranja šarenice ili retine mogu dobro funkcionirati u modu identifikacije, ali treba uzeti u obzir posljedice velikih baza podataka [1]. S druge strane, verifikacijski mod pruža veću fleksibilnost i obično zahtijeva manje resursa sustava, čime se postiže bolja učinkovitost. Zanimljivo je spomenuti i činjenicu da je u nekim praktičnim biometrijskim sustavima identifikacija polu-automatizirana [3]. U tom

slučaju sustav izdaje identitete top t podudaranja, a ljudski stručnjak ručno određuje identitet koji najbolje odgovara danom upitu. Dakle, za uspješnu implementaciju jednog od ova dva sustava ključno je temeljno planiranje, uzimanje svih relevantnih čimbenika, dobro upravljanje predlošcima, tehničke mogućnosti te očekivanja korisnika.[2]

3.4. Sigurnost biometrijskih sustava

U informacijskom svijetu, gdje se informacije dijele svake sekunde te postoji ogroman promet podataka, sigurnost postaje ključna značajka. Naravno, ništa drugačije nije niti s biometrijom. Biometrijski sustav bi trebao pružati zaštitu informacija od neautoriziranog pristupa, korištenja, otkrivanja, prekida, izmjene ili uništenja. Uobičajeno je da se u kontekstu sigurnosti informacija razmatraju četiri ključna aspekta [3]:

1. Integritet - zaštita od neprimjerenih izmjena ili uništenja podataka te osiguravanje neopozivosti i autentičnosti informacija.
2. Povjerljivost podataka - sprječavanje nelegitimnog pristupa ili otkrivanja osjetljivih informacija.
3. Dostupnost - jamčenje pravovremenog i pouzdanog pristupa informacijama i njihova korištenja.
4. Autentifikacija - omogućavanje pristupa podacima i izvođenje specifičnih zadataka samo legitimnim i autoriziranim korisnicima.

Općenito je poznato kako biometrijsko prepoznavanje može učinkovito riješiti problem autentifikacije. S obzirom na to da se biometrijska obilježja ne mogu lako izgubiti, ukrasti, zaboraviti ili dijeliti, biometrijsko prepoznavanje nudi elegantnije i pouzdanije rješenje za autentifikaciju u usporedbi s drugim tehnikama. Upravo zato se biometrijski sustavi sve više koriste za kontrolu pristupa u drugim informacijskim sustavima. Korištenjem biometrijskog sustava kao podsustava vladinih i ne vladinih organizacija moguće je pratiti sve korake i procese, odnosno tok povjerljivih informacija. Time se povećava odgovornost transakcija te ukida neovlašteni pristup povlaštenim informacijama. Iako takav sustav zadovoljava samo jedan aspekt u kontekstu sigurnosti informacija, njegovo kompromitiranje može bez problema ugroziti cijeli informacijski sustav, jer je zaobiđen proces autentifikacije. Ostale tehnologije poput enkripcije, digitalnog potpisa i sl. potrebne su za ispunjavanje aspekata za povjerljivošću, integritetom i dostupnošću ukupnog informacijskog sustava. Zbog toga je analiza sigurnosnih aspekata uključenih u dizajn i implementaciju biometrijskog sustava ključna i treba biti pažljivo i neovisno provedena.

Neautorizirano otkrivanje osjetljivih ili povjerljivih osobnih informacija može uzrokovati objektivne i subjektivne štete. U kategoriju objektivne štete ulazi financijska prevara ili uskraćivanje usluge, dok bi informacija o privatnim osobnim podacima pojedinca od strane druge osobe bila subjektivna šteta, čime se narušava ili gubi privatnost [1]. Iako je potreba za privatnošću obično individualna preferencija, postoje slučajevi kada može biti potrebno otkrivanje informacija u korist većeg društvenog interesa, primjerice kada se radi o nacionalnoj sigurnosti. Stoga biometrijsko prepoznavanje može poslužiti kao alat za očuvanje privatnosti pojedinaca ograničavanjem pristupa njihovim osobnim informacijama, međutim upotreba biometrije može sama po sebi stvoriti dilemu privatnosti. To je zato što biometrijski identifikatori pružaju neoborivu poveznicu prema identitetu osobe. Iz tog razloga, korisnici biometrijskih sustava imaju niz legitimnih zabrinutosti. Kad govorimo iz aspekta sigurnosti, postoji problem kod osoba koje su pod zaštitom identiteta, jer iako se vode pod drugim imenima i dalje posjeduju iste biometrijske karakteristike. Neke od mogućih dilema su i “Hoće li zahtjev za biometrijom biti proporcionalan potrebi za sigurnošću?“ te „Tko je vlasnik biometrijskih podataka, pojedinac ili pružatelji usluga?“. Takva pitanja o privatnosti su složena, no jedno je sigurno – nemaju konkretne odgovore. Principi pravedne prakse informacija kao što su transparentnost, informirani pristanak, ograničenje upotrebe, odgovornost i revizija mogli bi se slijediti kako bi se ograničila pitanja privatnosti koja proizlaze iz biometrijskog prepoznavanja. Budući da su ova pitanja izvan dosega tehnologije, potrebno je konzultirati se s odgovarajućim zakonodavstvenim tijelom, u svrhu provođenja principa. [3]

4. Biometrija glasa

Nakon općih pojmova o biometriji te njenim karakteristikama dolazimo i do glavne teme ovog diplomskog rada – biometrije glasa. Biometrija glasa je tehnologija koja omogućuje prepoznavanje i autentifikaciju korisnika na temelju njihovog glasa. U zadnjih par godina porastao je interes za internet stvari, a posebno za potkategoriju pametnih kuća, gdje ključnu ulogu igra upravo prepoznavanje glasa/govora te identifikacija glasa [5]. Stavljajući biometriju glasa na prvo mjesto pristupačnosti biometrijskih osobina, jer nisu potrebni nikakvi dodatni uređaji ili prijenosni sustavi [1]. Opće je poznato da je ljudski glas jedinstven za svaku osobu, konstruirajući specifičan uzorak frekvencija i karakteristika. Prema tome ponovno možemo podijeliti sustave biometrije glasa na identifikacijske i verifikacijske. Osim te podjele postoje još dvije kategorizacije unutar industrija - procesiranje glasa i biometrijska sigurnost. [6]

4.1. Procesiranje glasa i biometrijska sigurnost

Za procesiranje glasa, kao i kod drugih alata za procesiranje glasa, biometrijski sustavi glasa ekstrahiraju podatke iz govornog toka kako bi uspješno obavili namijenjeni posao. Za obavljanje posla, biometrijski sustav moguće je konfigurirati s mnogim istim akustičnim parametrima kao i kod srodne grane - prepoznavanje govora. Zato i ne čudi da upravo između tih dvaju pojmova dolazi do zabune i pogrešne interpretacije. Također, kao što je to slučaj s prepoznavanjem govora, alati za procesiranje glasa imaju koristi od velike količine podataka, kvalitetnih mikrofona i softvera za poništavanje buke. Biometrijski sustavi glasa su osjetljivi na neke od istih uvjeta koji uzrokuju loše performanse sustava za prepoznavanje govora [6]: pozadinska i kanalna buka, promjenjivi i inferiorni mikrofoni i telefoni te ekstremna promuklost, umor ili stres glasa. Što nužno ne mora biti loše, kako ćemo vidjeti u sljedećim poglavljima. Postoje važne razlike između biometrijskih sustava glasa i drugih tehnologija za obradu govora, uključujući prepoznavanje govora. Najznačajnija je ta što tehnologije biometrijskog prepoznavanja glasa ne znaju što osoba govori, oslanjajući se na prepoznavanje govora da to učini. Štoviše, trend prema neovisnosti o govorniku, koji karakterizira prepoznavanje govora, ne može postojati za biometrijsko prepoznavanje glasa. Po definiciji, biometrijsko prepoznavanje glasa uvijek je povezano s određenim govornikom [1]. Kao rezultat, oni zahtijevaju neku vrstu registracije za svakog korisnika. Potreba za registracijom je značajka koju biometrijsko prepoznavanje glasa dijeli sa svojim srodnicima u industriji biometrijske sigurnosti. Tehnologije zasnovane na biometriji najčešće se primjenjuju u sigurnosti, nadzoru i prevenciji prijevara gdje pozitivno identificiraju pojedince i razlikuju jednu osobu od druge. Te sposobnosti razlikuju biometriju od svih drugih oblika automatizirane sigurnosti. Kako je već

prije navedeno sustav s karticama može, u najboljem slučaju, utvrditi samo posjeduje li osoba valjanu pristupnu karticu, a sigurnost lozinke može utvrditi samo zna li osoba ispravnu lozinku. Nijedan od njih ne provjerava je li osoba koja predstavlja karticu ili unosi lozinku ovlaštena za to. Biometrijski sustavi utvrđuju dolazi li biometrijski uzorak, poput otiska prsta ili izgovorene lozinke, od određene osobe uspoređujući taj uzorak s referentnim biometrijskim uzorkom – uzorkom iste vrste biometrije koji je pružila dotična osoba. Programeri biometrije prepoznavanja glasa nazivaju to "referentnim glasovnim otiskom". Kao i kod referentnih predložaka za druge biometrije, referentni glasovni otisci procjenjuju se prema broju puta kada pogrešno prihvate lažni zahtjev za identitetom kao legitimni zahtjev i broju puta kada odbiju legitimnog govornika kao prevaranta. [6]

Najznačajnija, a ujedno i najzanimljivija razlika između biometrijskog prepoznavanja glasa i ostalih biometrija je ta što su biometrijski sustavi glasa jedini komercijalni biometrijski sustavi koji koriste i procesiraju akustične informacije. Većina ostalih biometrija temelji se na slikama. Još jedna važna razlika je ta što su većina komercijalnih biometrijskih sustava glasa dizajnirana za korištenje s gotovo bilo kojim standardnim telefonom na javnim telefonskim mrežama. Mogućnost rada sa standardnom telefonskom opremom omogućava široku primjenu biometrijskih aplikacija glasa u različitim postavkama. Dok s druge strane, većina ostalih biometrija zahtijeva vlastitu hardversku opremu, poput senzora otiska prsta ili opreme za skeniranje šarenice. [6]

4.2. Fiziologija govornog sustava i način nastanka glasa

Kako bi što bolje razumjeli način na koji funkcionira biometrija glasa, nije na odmet upoznati se ukratko i s načinom nastanka glasa, odnosno same fiziologije govornog sustava. Fiziologija govornog sustava i način nastanka glasa su ključne komponente u razumijevanju kako ljudi proizvode govor. Ovaj proces je kompleksan i uključuje nekoliko različitih dijelova tijela koji rade zajedno kako bi omogućili kreiranje i artikulaciju zvuka.

4.2.1. Anatomija i fiziologija govornog sustava

Govorni sustav sastoji se od nekoliko ključnih komponenti, uključujući respiratorni sustav (pluća), fonatorni sustav (glasnice), i artikulacijski sustav (jezik, zubi, nepce). Sve ove komponente igraju ključnu ulogu u nastanku glasa. Respiratorni sustav pruža aerodinamičku energiju potrebnu za nastanak glasa. Zrak se iz pluća propušta kroz traheju, pokrećući vibraciju

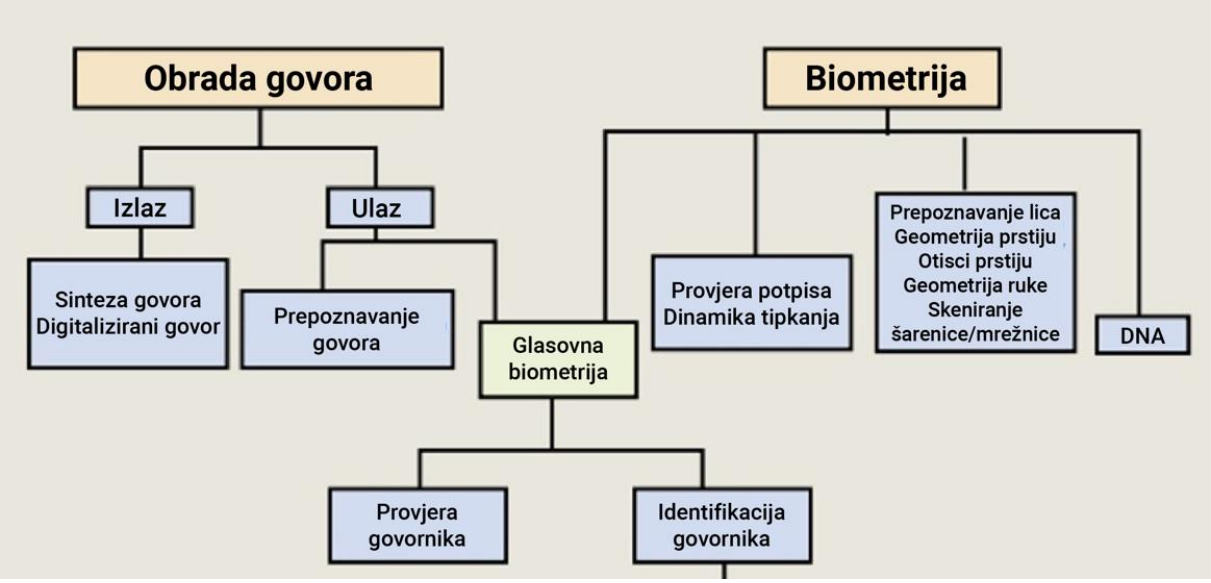
glasnica. Fonatorni sustav koristi vibracije glasnica za stvaranje zvukova. Dok artikulacijski sustav oblikuje i modificira osnovni zvuk koji proizvode glasnice u razumljive glasove i riječi. Kretanjem i pozicioniranjem ovih dijelova tijela, govor se artikulira i modulira pritom formirajući različite glasove i tonove. Dakle nastanak glasa započinje disanjem, gdje zrak ulazi u pluća i zatim se izbacuje prema gore kroz respiratorni trakt. Ključan trenutak u proizvodnji glasa događa se u glasnicama. Ovisno o njihovoj napetosti, koja se može kontrolirati mišićima, brzini protoka zraka i njihovom položaju, vibracije glasnica mogu proizvesti različite visine zvuka. Za artikulaciju glasa, artikulacijski organi (jezik, usne i nepce) mijenjaju oblik i položaj kako bi modificirali zvukove koji dolaze iz glasnica. Ove promjene u zračnom toku i obliku usne šupljine proizvode specifične glasove i riječi.[1]

4.2.2. Karakteristike glasa koje se koriste u biometriji

Dosad smo vidjeli da je nastajanje glasa vrlo kompleksno i da bi se o njemu moglo pisati u detalje. No nas u biometriji glasa zanima dio finalnog glasovnog signala kojeg je generirao pojedinac. U obje faze proizvodnje glasa (jezično generiranje i proizvodnja govora) uvode se specifičnosti govornika [1]. Unutar polja biometrije glasa te dvije karakteristike možemo podijeliti na dvije razine: visoki nivo (lingvistika) i niski nivo (akustika). Mnogobrojna istraživanja su potvrdila da ljudi prepoznaju govornika prema kombinaciji različitih informacija pritom koristeći različite težine za dobivene informacije. Na primjer visina glasa, nazalnost i slično. Nadalje postoje i idiolektalne karakteristike govornika, koje su na najvišoj razini koju tehnologija do sada obično uzima u obzir. One opisuju kako govornik koristi specifični jezični sustav. Ova "upotreba" određena je mnogobrojnim faktorima, neki od njih su prilično stabilni kod odraslih, kao što su razina obrazovanja, sociološki i obiteljski uvjeti te podrijetlo. Međutim, postoje i neki faktori visokog nivoa koji su uvelike ovisni o okolini. Tako primjerice doktor ne koristi jezik na isti način kada razgovara sa svojim kolegama u bolnici, sa svojom obitelji kod kuće ili s prijateljima. Za idiolektalno prepoznavanje govornika, biometrijski sustavi glasa koriste razne algoritme i baze podataka za usporedbe različitih jezičnih uzoraka. U drugu grupu karakteristika koje se koriste u biometriji spada fonotaktika, koja je bitna za ispravnu upotrebu jezika i igra važnu ulogu prilikom učenja stranih jezika. Osim nje postoje je prozodijske karakteristike koje uključuju energiju izgovaranja, intonacija te brzinu govora i trajanja pauza u govoru [5]. Zbog toga govor može biti prirodan, imati smisao i emotivan ton. Ukratko prozodija pomaže pri razumijevanju generirane poruke govornika. Na kraju na nižem nivou postoje spektralne karakteristike govornih signala, koje su usko povezane s govornikom i njegovim artikulacijskim radnjama povezanima sa svakim fonemom koji se proizvodi, kao i s pojedinačnom fiziološkom konfiguracijom aparata za proizvodnju govora.

4.3. Tipovi biometrije glasa

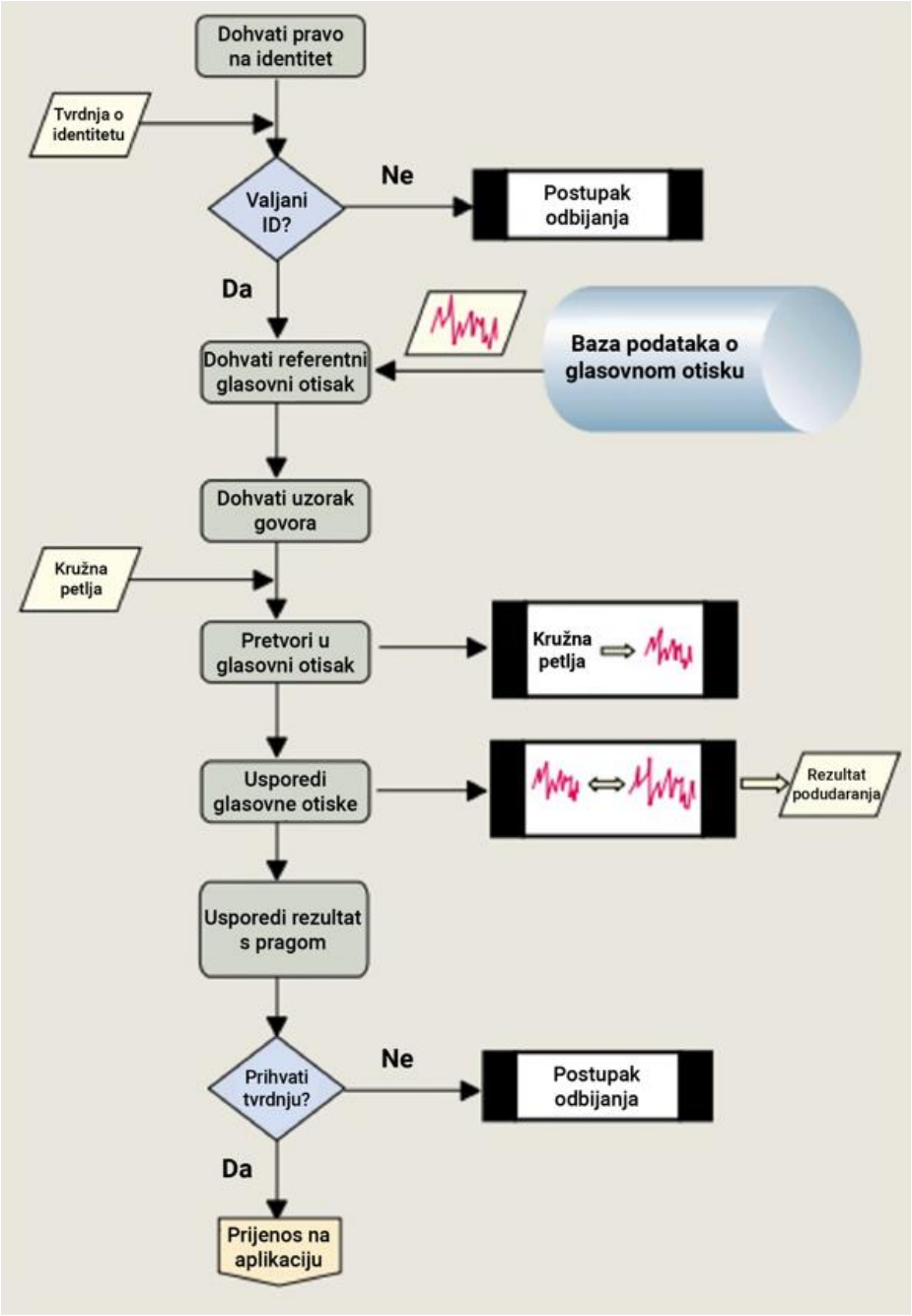
U ovoj cjelini bit će detaljnije opisana podjela biometrije glasa na identifikacijsku i verifikacijsku razinu. Na sljedećoj slici možemo vidjeti povezanosti biometrije, procesiranja govora i biometrije glasa.



Slika 3. Prikaz rodnog stabla biometrije glasa. [6]

4.3.1. Verifikacija govornika

Sustavi verifikacije govornika autentificiraju osobu potvrđujući njezin identitet na temelju njezinog glasa. Na primjer, ako bi osoba uzviknula "Ja sam Ana" umjesto "To sam ja!", namijenjeni slušatelj morao bi izvršiti verifikaciju govornika na temelju izrečene tvrdnje. Isto vrijedi i za sustav verifikacije. Proces verifikacije u takvim sustavima započinje tvrdnjom identiteta, prisjetimo se to je referentna oznaka osobe koja je spremljena unutar baze sustava [1]. Moguće je da takvi sustavi koriste i prepoznavanje glasa, tada će se od osobe tražiti da izgovori svoj referentni broj. Sljedeći korak je pronalazak glasovnog otiska iz baze te traženje uzorka od podnositelja zahtjeva. Novo uneseni govor pretvara se u glasovni otisak i uspoređuje s referentnim glasovnim otiskom. Rezultati usporedbe kvantificiraju se i uspoređuju s pragom prihvaćanja/odbijanja kako bi se utvrdilo jesu li dva glasovna otiska dovoljno slična da sustav prihvati tvrdnju o identitetu. U nastavku je ilustracija pseudo procesa, opisanih koraka verifikacije:



Slika 4. Prikaz procesa verifikacije govornika. [6]

Postoje različiti načini interakcije sa sustavima verifikacije govornika. Većina komercijalnih sustava ovisi o tekstu. Traže lozinku, broj računa ili neki drugi unaprijed dogovoreni kod. Budući da traže lozinku, sustav je ovisan o tekstu. Ovisni sustavi o tekstu pružaju tzv. "snažnu autentifikaciju", koja zahtijeva upotrebu barem dvije različite vrste sigurnosti. U slučaju verifikacije govornika ovisne o tekstu, osoba mora imati ispravan uzorak glasa (primjer sigurnosti "Tko ste") i znati ispravnu lozinku (primjer sigurnosti "Što znate"). Druga vrsta interakcije sa sustavima je korištenje isključivo glasa, tada prepoznavanje govora dekodira unos, a verifikacija govornika koristi isti unos kao biometrijski uzorak koji se

uspoređuje s referentnim glasovnim otiskom. Postoje i "tekstom potaknuti" sustavi koji traže od govornika da ponovi niz nasumično odabranih nizova brojeva, riječi ili fraza. Tekstom potaknuta tehnologija zahtijeva duže vrijeme upisa jer generirani referentni glasovni otisak mora sadržavati sve komponente koje će se koristiti za konstruiranje varijanti izazova-odgovora. Verifikacija u tom slučaju također traje duže. Zadnja vrsta su sustavi čija verifikacija ne ovisi o tekstu. U takvim slučajevima nema nikakvih potvrda, već korisnik razgovara s agentom, a sustav u međuvremenu verificira da se radi o osobi za koju se korisnik predstavio.

Iako je „tekstom potaknuta“ verifikacija značajno sigurnija, u današnje vrijeme ne pronalazi široku primjenu jer je prilično teško pobijediti dobar komercijalni sustav verifikacije govornika snimkom. Glasovni signal unesen u mikrofon ili telefon držan blizu usta govornika znatno se razlikuje od signala snimljenog čak i na udaljenosti od metar ili dva. Osim toga, mnogi komercijalni sustavi verifikacije govornika traže znakove koji upućuju na upotrebu snimke. Kao rezultat, stvaranje snimke koja može prevariti sposobnost aplikacije za rad u pozadini daje tehnologiji neovisnoj o tekstu potencijal za korištenje bez znanja subjekta.[6]

4.3.2. Identifikacija govornika

Za razliku od verifikacije, identifikacija govornika proces je dodjeljivanja identiteta glasu nepoznatog govornika. Dakle verifikacija govornika, potvrđuje identitet na temelju prethodne tvrdnje, a identifikacija zahtjeva prepoznavanje govornika isključivo na temelju njegovog glasa [1]. Ovaj proces obično je složeniji jer uključuje usporedbu više govornih uzoraka koji se međusobno mogu znatno razlikovati i mogu biti snimljeni različitom opremom. U forenzičkim i obavještajnim primjenama, uzorci mogu biti prikupljeni skrivenom opremom u bučnom okruženju, što dodatno otežava identifikaciju. Proces identifikacije govornika započinje kada se uzorak govora nepoznatog govornika predstavi sustavu, zatim započinje obrada koja uključuje pretvorbu govora u glasovni otisak te usporedba primljenog otiska s referentnim glasovnim otiscima u bazi podataka. Sustavi mogu biti konfigurirani da rangiraju referentne glasovne otiske prema vjerojatnosti da sadrže glas osobe koja generira uzorak, ili da izaberu jedan ili dva potencijalna identiteta. Postoje i implementacije koje omogućavaju sustavu da prijavi neusklađenost glasa s bilo kojim referentnim glasovnim otiskom, koristeći se strategijom za kontrolu prijevare s mobilnim pozivima. Za poboljšanje performansi sustava, uobičajene metode uključuju prilagodbu glasovnog otiska, modeliranje kohorte i svjetske modele. Prilagodba glasovnog otiska uključuje ažuriranje originalnog glasovnog otiska s podacima iz uspješnih verifikacija, dok modeliranje kohorte i svjetski modeli predstavljaju tehnike poboljšanja performansi kroz identifikaciju sličnih glasova ili usporedbu s općim modelom koji sadrži širok spektar glasova. Temeljna pretpostavka je da će novi ulaz od ovlaštenih korisnika

bolje odgovarati njihovim referentnim glasovnim otiscima nego otiscima drugih osoba, čak i u nepovoljnim uvjetima.[6]

4.3.3. Koraci analize glasa i optimizacije biometrijskih sustava glasa

U ovoj cjelini ukratko ćemo proći kroz korake za prepoznavanje govornika, odnosno što je sve potrebno za dobru glasovnu biometriju, te sumirati dosadašnje činjenice vezane za biometriju glasa.

Proces je prilično jednostavan, zabilježi se uzorak glasa koji se zatim analizira kako bi se dobile karakteristike potrebne za identifikaciju osobe. Ritam, visina tona, frekvencija i boja glasa neke su od karakteristika koje se koriste u analizi glasa. Sustavi biometrije glasa koriste algoritme strojnog učenja i baze podataka uzoraka glasa. Zatim algoritmi analiziraju i uspoređuju poznate karakteristike glasa s nepoznatima. Ako je podudaranje glasova dovoljno visoko, sustav može s pouzdanošću odrediti tko govori. Za kvalitetan biometrijski uzorak glasa ključna je pravilna obuka modela s obiljem uzoraka glasa tijekom registracije. Smatra se da biometrija glasa može izazvati sumnje zbog promjena glasa uslijed bolesti, umora ili s vremenom. Ipak, upotreba tehnologije dubokog učenja i stručnost inženjera mogu prevladati sve izazove i identificirati glas korisnika u različitim scenarijima. Najčešći slučajevi uključuju verifikaciju identiteta u sigurnosnim sustavima i autentifikaciju korisnika u pozivnim centrima. U nastavku su navedeni koraci analize glasa i značajke dobre prakse. [1]

1. **Ekstrakcija značajki i tokenizacija** - Prvi korak u izgradnji sustava za automatsko prepoznavanje govornika je pouzdano izdvajanje značajki i tokena koji sadrže identificirajuće informacije govornika. Ekstrakcija uključuje niži nivo značajki (spektralne informacije, energija glasa, visina tona) te srednji i viši nivo tokena kao što su fonemi, slogovi i riječi.

2. **Kratkoročna analiza** - Radi pouzdane spektralne analize, signali moraju pokazivati stacionarna svojstva koja nisu lako uočljiva u stalno mijenjajućim govorom signalima. Analiza se ograničava na kratke dužine između 20 i 40 ms kako bi se dobili pseudo-stacionarni signali po okviru.

3. **Parametrizacija** – korištenje „hamming/hanning“ okvira dobivamo sadržaj željenih vremenskih/spektralnih informacija. Linearna prediktivna kodiranja (LPC) govora pokazala su se kao valjan način za kompresiju spektralne omotnice u modelu sa svim polovima sa samo 10 do 16 koeficijenata.

4. **Fonetička i riječna tokenizacija** - Skriveni Markovljevi modeli (HMM) najuspješniji su i najšire korišteni alat za fonetičku, slogovnu i riječnu tokenizaciju, tj. prijevod uzorkovanog govora u vremenski poravnani niz jezičnih jedinica.

5. **Prozodijska tokenizacija** - Osnovne prozodijske značajke kao što su visina tona i energija također se dobivaju na razini okvira. Energijski okvir vrlo je lako dobiti kroz Parsevalov teorem, a trenutačna visina tona može se odrediti metodama kao što su auto korelacija ili kepstralna dekompozicija.

Uz navedene korake analize, nije na odmet nakon analize napraviti generalizaciju podataka, odnosno klasificirati sustave i tehnike. Možemo klasificirati sustave prepoznavanja govornika ovisne o tekstu iz perspektive primjene na sustave s fiksnim tekstom i sustave s promjenjivim tekstom [6]. Također, ako postoje već neke baze podataka, kod modernijih pristupa otkrivanja emocija iz snimke glasa, koje koriste neuronske mreže, tada ih možemo koristiti u treniranju modela. U takvim slučajevima dobro istrenirani modeli mogu značajno poboljšati rezultate prepoznavanja emocija.

4.4. Primjene biometrije glasa i njezina važnost

Biometrija glasa, kao tehnologija koja koristi jedinstvene karakteristike glasa pojedinca za identifikaciju, nalazi primjenu u različitim sektorima, uključujući bankarstvo, sigurnost, telekomunikacije i zdravstvenu skrb. Korištenje glasovne biometrije omogućava brzu i sigurnu autentifikaciju korisnika, smanjuje rizik od prijevare i pruža jednostavno korisničko iskustvo. Osim poboljšanja sigurnosti, primjene biometrije glasa pružaju značajne prednosti u pogledu dostupnosti i pristupačnosti, omogućujući korisnicima da obavljaju različite zadatke poput plaćanja, pristupa uslugama ili zaključavanja i otključavanja uređaja svojim glasom. Evo i nekoliko detaljnih primjena upotrebe i njezine važnosti za pojedini slučaj:

1. **Komercijalne svrhe** - Autentifikacija u različite komercijalne svrhe od otključavanja telefona, do brzog potvrđivanja identiteta za različite osobne usluge.

2. **Forenzika** - Ako je uzorak govora snimljen tijekom izvršenja krivičnog djela, može se usporediti s glasom osumnjičenog da bi se pronašla bilo kakva sličnost između njih. U ovom primjeru vidimo važnost biometrije glasa za očuvanje sigurnosti građana.[6]

3. **Fintech** - Mnoge banke su dodale glasovnu biometriju kao sekundarni nivo autentifikacije korisnika. [6]

4. **Kontrola pristupa** - Limitirani pristup do baza podataka ili kritičnih informacija, primjerice u vojsci. [6]

5. **Preventivna sigurnost** - Analiza razgovora između terorista može utvrditi odstupanja od klasičnih razgovora, te tako označiti komunikaciju za daljnju detaljniju analizu.

6. **Dodatna sigurnost** - Postoje one tvrtke koje žele dodatno povećati sigurnost ili pristup određenim podacima. Prepoznavanje glasa dodaje još jedan sloj sigurnosti, odnosno sekundarna pouzdana metoda autentifikacije.

7. **Zdravstvo** - Od verifikacije pacijenata do psihologije i analiziranja emocija pacijenata. Nekada se na prvu može zaključiti da je sve u redu, međutim detaljnom analizom i jakim biometrijskim modelima za prepoznavanje emocija iz snimke glasa dobijemo detaljniji uvid u emotivno stanje osobe.[6]

8. **Osobna asistencija** - Virtualni asistenti na mobilnim uređajima ili uređaji za upravljanje pametnim domovima kao što je Amazon Alexa koriste biometrije glasa za izvršavanje naredbi ili kupovinu. Alexa može prepoznati tko joj se od ukućana obraća te prema tome zna koja ovlaštenja imaju. Primjerice dijete ne može izvršiti transakciju kupovine. Osim toga uređaji bazirani na prepoznavanju osjećaja mogu lakše odgovoriti osobi na temelju analize njenih osjećaja te ovisno o njenom trenutnom stanju formirati prikladniji odgovor.[5]

9. **Glasovne pretrage** - Razne tražilice imaju integrirane mehanizme za prepoznavanje govora, poznatiji kao i speech-to-text.

10. **Marketing** - Analizom govora korisnika moguće je prikazivati oglase ovisno o raspoloženju ili ih učiniti personaliziranijima te prilagoditi sučelja na temelju njihovih preferencija i zahtjeva. [5]

11. **Pozivni centri** - Najučestaliji korisnici biometrije glasa, omogućuju brzu potvrdu identiteta te ubrzavaju rad.[7]

12. **Alat za detekciju prijevara** - Sustavi biometrije glasa mogu služiti i kao moćan alat za detekciju prijevara. Zbog sve većih prijevara i krađa identiteta državna tijela koriste biometriju glasa u te svrhe.[7]

13. **Digitalno potpisivanje** - S obzirom na to da je glas osobe jedinstven, moguće je digitalno potpisivanje dokumenata putem svog glasa. Što otvara veliki prostor za unapređenje u firmama kod potpisivanja ugovora.[7]

14. **Upravljanje zaposlenicima** - Upotreba glasovne biometrije u aplikacijama za upravljanje zaposlenicima trenutno je manje uobičajena, no može služiti evidenciji dolazaka i odlazaka radnika ili ubrzati proces identifikacije za razliku od biometrije otiska prsta. [7]

15. **Transkripcija i kontrola uređaja** - Sustavi za dikciju i kontrolu uređaja često koriste prepoznavanje govora kako bi omogućili ljudima diktiranje teksta ili kontrolu uređaja s pomoću glasovnih naredbi. [5]

16. **Sigurnost putovanja** - Neki aerodromi i kompanije za mobilnost koriste glasovnu biometriju za provjeru identiteta putnika i zaposlenika. Osigurava da samo ovlaštene osobe uđu u avione, vlakove ili autobuse. [5]

4.5. Prednosti i ograničenja biometrije glasa

Korištenje biometrije glasa i srodnih tehnologija kao prepoznavanje glasa omogućuje multi-tasking i udobnost bez upotrebe ruku [5]. Iako se tehnologija prepoznavanja glasa ubrzano poboljšava, nije u potpunosti bez grešaka. Razgovaranje i davanje glasovnih komandi je mnogo brže od unosa putem tipkovnice. Pozadina buke može ometati rad i uticati na pouzdanost biometrijskog sustava. Slučajevi upotrebe prepoznavanja glasa se šire sa strojnim učenjem kao i dubokim neuronskim mrežama. Bitno je u takvim situacijama staviti i naglasak na sigurnost i privatnost snimljenih podataka.

Primjena biometrije glasa posebno se ističe zbog svoje sposobnosti da funkcionira bez posebnih čitača i na udaljenim lokacijama, poput telefonskih transakcija. Međutim, uporaba ove tehnologije ima i svoja ograničenja, u nastavku slijedi popis prednosti i nedostataka korištenja biometrije glasa:

1. **Prednosti:**

    a. **Niski operativni troškovi** - Veliki sustavi kao bankarski ili pozivni centri mogu uštedjeti puno korištenjem biometrijskih sustava glasa. Ušteda se evidentira kroz smanjivanje potrebnih koraka koje koriste ostale verifikacijske metode. Također moguće je identificirati ili verificirati korisnike za vrijeme razgovora, bez potreba za personaliziranim upitima. [7]

    b. **Poboljšano korisničko iskustvo** - Nema potrebe za unašanjem raznih pinova ili odgovaranja na izazovna pitanja (nešto slično kao reCAPTCHA). Pružajući više kanalnu komunikaciju prema svim servisima tvrtke (slično kao SSO na webu). [7]

    c. **Povećana preciznost** - Autentifikacija glasom je preciznija i vjerodostojnija od korištenja lozinki i pinova, koje je moguće zaboraviti, ili ako se radi o brute force napadu pogoditi. Usprkos činjenici da zvučni zapis može biti narušen raznim stvarima i dalje je daleko praktičniji jer smanjuje prostor za greške i povećava produktivnost. [7]

    d. **Laka implementacija** - Vrlo bitna stvar kod dizajniranja i odabira biometrijskih sustava je način i težina implementiranja u postojeće sustave. Mnoge tvrtke vrednuju upravo jednostavnost korištenja i vrijeme implementacije/nadogradnje sustava. Kako je već spomenuto za korištenje glasa nije potreban nikakav dodatni hardware, naravno osim mikrofona. Zbog toga je moguće dio zaposlenika realocirati na druge poslove, smanjujući potrebu za ljudskim resursima. [7]

    e. **Kvazi-stacionarnost** - Govorni signal je kvazi-stacionaran kada se analizira u kratkim vremenskim intervalima. To omogućava ekstrakciju značajki s visokom diskriminativnom moći za razlikovanje među pojedincima.[5]

    f. **Niska razina nametljivosti** - jednostavan za korištenje, hands-free model.[5]

    g. **Dostupnost** - cijena je jedan od ključnih faktora za tvrtke prilikom odabira biometrijskog sustava. Snimanje glasovne poruke smatra se standardnom

metodom identifikacije. Ljudi su navikli na ovakav način metode. Ne smatraju se neobičnima.[5]

h. **Pristupačnost** - prednost za osobe s invaliditetom, koje imaju problema s pisanjem ili korištenjem taktilnih senzora.[5]

2. Ograničenja [5]

a. **Utjecaj okoline** - jedan od izazova s kojima se suočava je utjecaj vanjskih zvukova, koji može povećati stopu pogrešne odbijenosti do 2 %.

b. **Baze podataka** - podatci na kojima su trenirani biometrijski sustavi glasa mogu davati različite odgovore ovisno o spolu, jeziku ili rasi. Zato je potrebno trenirati modele s audio zapisima koji su čim više heterogeni.

c. **Naglasak** - ovakva tehnologija može biti manje učinkovita za ljude koji imaju drugačiji naglasak jer sustav u takvim situacijama ima poteškoće u prepoznavanju izgovorenih riječi ili fraza.

d. **Izgovaranje riječi** - slično kao i kod naglaska sustav može biti manje precizan za osobe koje imaju problema prilikom izgovaranja riječi (mucanje) ili koji mumljaju dok pričaju.

e. **Nedostatak podataka** - u slučaju da sustav nema dovoljno podataka za skup osoba koje treba identificirati ili verificirati, očekivana je manja stopa preciznosti.

Možemo primijetiti kako je ipak više prednosti nego ograničenja biometrije glasa, a uz pravilnu optimizaciju i tehnološko znanje većina prepreka se može zaobići.

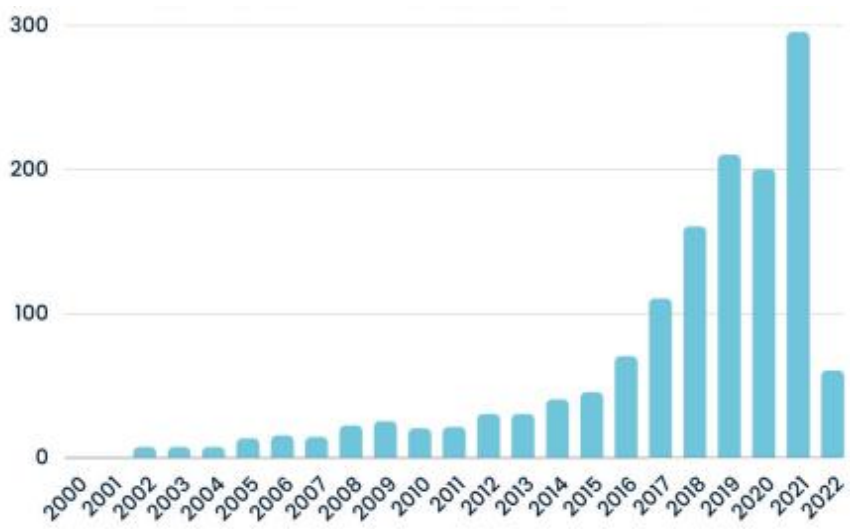
4.6. Budućnost biometrije glasa

Biometrija glasa je od početka tisućljeća dobila na popularnosti i interesu u daljnja istraživanja u tom području. Od 2000. do danas rast je eksponencijalan, međutim tu nije kraj. Predviđa se da bi tržište biometrije glasa i prepoznavanja glasa, s trenutnih 1.1 bilijuna dolara (2020.) moglo narasti do skoro 4 bilijuna dolara u sljedećih dvije godine. To bi značilo rast od

22.8 % godišnje, što nije zanemarivo [7]. U korist idu i činjenice da nam manjka sigurnosnih značajki, međusobne interakcije uređaja kao primjerice u bankarskom sektoru, gdje bilježe veliki porast prevara i napada. Prema statistici, u nazad četiri godine, došlo je do povećanja za 269 %, što je mnogo više od ostalih industrija zajedno. Upravo zato ne čudi da bankovni sektor, kao i razne fintech kompanije implementiraju biometrije glasa i prepoznavanje glasa u svoje sustave [5], radi povećanja sigurnosti i smanjivanja postotka prevara. Ovakvu prilagodbu možemo vidjeti i kod naše najpoznatije fintech tvrtke Aircash, gdje prilikom registracije aplikacija od korisnika traži da na glas pročita brojeve s ekrana, vjerojatno oslanjajući se na jednu od varijanti verifikacije - ovisnosti o tekstu uz prepoznavanje govora ili tekstom potaknutu verifikaciju. Uz to koriste vjerojatno i neku od biometrija koja prepoznaje lice korisnika. U ovom primjeru možemo vidjeti na djelu multibiometriju, koja omogućuje dodatnu sigurnost. Dakle, ove tehnologije donose veliku transformaciju raznih postojećih sustava, nudeći sigurniju i prikladniju autentifikaciju za korisnike. S razvojem mobilnih aplikacija i napretkom u tehnologiji prepoznavanja glasa, očekuje se da će glas postati još važnije sredstvo biometrijske identifikacije u budućnosti.

5. Prepoznavanje emocija iz snimke glasa

Govor sadrži niz značajki koje slušatelj interpretira kako bi dešifrirao bogate informacije koje mu govornik želi prenijeti. Govornik također nenamjerno dijeli razne akustičke karakteristike glasa kao što su ton, energija, brzina te druge akustičke osobine, što omogućuje prepoznavanje konteksta i davanje značaja izgovorenom sadržaju. Rad na prepoznavanju govora započeo je pretvaranjem govora u tekst odnosno stvaranjem transkripta. Time je zabilježena prva razina informacija točnije značenje govora. U naprednijim primjenama, kontekst i suosjećanje s govornikom postaju ključni za prepoznavanje emocija u govoru. Ovdje se također vidi razlika između analize sentimenta teksta i prepoznavanja emocija u govoru. U analizi sentimenta, emocija se doslovno prenosi u tekstu, što olakšava razumijevanje dobivene poruke (sretno, ljuto, tužno…). Međutim, u prepoznavanju emocija u govoru (eng. Speech Emotion Recognition, skraćeno SER), sve te informacije skrivene su ispod prvog sloja informacija. [8]



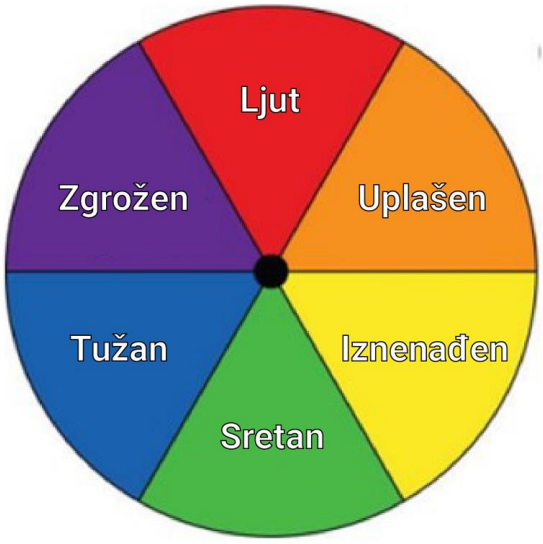
Slika 5. Broj publikacija na temu strojnog učenja i prepoznavanja emocija kroz godine. [8]

Na slici iznad vidljivo je kako interes za ovo područje raste eksponencijalno kroz godine. To je zbog napretka u algoritmima i primjene u stvarnim scenarijima. Ljudski govor sadrži paralingvističke informacije koje se mogu predstaviti kvantitativnim značajkama poput visine tona, intenziteta i Mel-frekvencijskih kepstralnih koeficijenata (eng. MFCC) [1]. SER se obično postiže kroz tri ključna koraka: obrada podataka, odabir/izdvajanje značajki i klasifikacija na temelju emocionalnih značajki. Priroda ovih koraka, zajedno s posebnim značajkama ljudskog govora, podržava korištenje metoda strojnog učenja za implementaciju SER-a.

5.1. Što su emocije i kako ih klasificiramo?

Definicija i percepcija emocija općenito se smatraju poznatim znanjem, no zapravo su jedan od najsubjektivnijih aspekata ljudske interakcije i komunikacije. Subjektivnost, osobne i kulturne razlike mijenjaju način na koji ljudi izražavaju i tumače emocije. S obzirom na to da u psihologiji ne postoji opće prihvaćeni znanstveni konsenzus o taksonomiji i mjerenju emocija, dolazi do izazova s kojima se susreće u istraživanju detekcije emocija iz snimke glasa. Psihologija razlikuje osjećaje, emocije i afekte ovisno o intenzitetu, trajanju i postojanosti.[22]

U psihologiji postoji mnogo teorija i klasifikacija emocija. U nekim od teorija, emocije su hijerarhijski organizirane prema određenim kriterijima, dok u drugim, kojih je usput i manje, sve emocije se smatraju jednako važnima. Unatoč razlici između teorija u jednom se slažu sve teorije - kompleksnost emocionalnog ponašanja ljudi. Tako su neke od tih teorija identificirale i proučavale 65 različitih emocija, a neki psiholozi su istaknuli da su "ljudske emocije izrazito raznolike". Jedna od najpoznatijih klasifikacija emocija pripada američkom psihologu Robertu Plutchiku [9]. U njegovoj taksonomiji postoji 8 temeljnih emocija, grupiranih u 4 para suprotnosti. Sve se one manifestiraju u različitim stupnjevima intenziteta, a njihova kombinacija rezultira sekundarnim emocijama. Ova složena shema grafički je prikazana od strane Plutchika na vrlo sugestivan način, poznat od tada kao 'Plutchikov kotač'. Kako se može vidjeti u ovoj reprezentaciji (Slika 7), temeljne emocije u različitim stupnjevima intenziteta povezane su s velikim brojem emocija, a one prikazane u središnjem dijelu smatraju se manifestacijama ekstremne intenzitete tih temeljnih emocija.[9]



Slika 6. Prikaz pojednostavljenog Plutchikovog kotača. [9]

Slika 7. Plutchikovo kotač emocija. [9]

Dolazimo do zanimljivog pitanja - Kako računala i drugi strojevi mogu detektirati i prepoznati emocionalne informacije? Koriste pasivne senzore za prikupljanje podataka o fizičkom stanju ili ponašanju korisnika – video kamera može snimiti izraze lica, držanje tijela i geste, dok mikrofon može snimiti govor. Prepoznavanje emocionalnih informacija zahtijeva ekstrakciju značajnih obrazaca iz prikupljenih podataka. Međutim, uzimajući u obzir složenost ljudskih emocija, postoji opće suglasje o tome da su neke od njih važnije za komunikaciju i interakciju s računalima. Imajući to na umu dolazimo i do pojednostavljene verzije Plutchikovog kotača (slika 6), koja je referenca u automatskom prepoznavanju emocija.

Nije iznenađujuće da istraživanje za prepoznavanje emocija zahtijeva interdisciplinarno znanje, s uvidima iz područja psihologije i kognitivnih znanosti, računalnih znanosti i elektronike, te također medicine ovisno o vrsti prepoznavanja. Tipično, ljudske emocije mogu se detektirati iz prepoznavanja lica, govora/glasa, jezika tijela/gestikulacija, analize bio-signala. Metode prepoznavanja emocija na temelju glasa opravdane su činjenicom da ljudski glas može prenijeti širok spektar emocija: od radosti do boli, od tjeskobe do sreće, od spontanosti

do krutosti, od nježnosti do grubosti, od zdravlja do bolesti, od smijeha do plača. Znanstvena istraživanja pokazala su da emocije svake osobe pokreću neke psihološke i fiziološke promjene koje utječu na glas. Neki psiholozi su pokazali da većinu vremena ne osjećamo samo jednu emociju, već niz dviju ili više emocija. Mnoga znanstvena istraživanja iz područja umjetne inteligencije usmjerena su na izjednačavanje emocionalne snage ljudskog glasa. Od 1981. godine razvijeni su neki algoritmi za prepoznavanje emocija u ljudskom glasu. Najbolji algoritmi koriste konvolucijske neuronske mreže, o kojima će biti detaljnije u sljedećem poglavlju, za identificiranje specifičnog obrasca za svaku emociju. Neki od tih algoritama već su implementirani u različitim inteligentnim uređajima koji mogu interaktivno komunicirati s ljudima na "empatičan" način i pružiti im povratne informacije povezane s njihovim emocionalnim stanjima.[9]

5.2. Metode za prepoznavanje emocija iz snimke glasa

Nedavna istraživanja u afektivnom računarstvu koristila su različite metode strojnog učenja (eng. Machine learning, skraćeno ML) za zadatke SER-a međutim, samo nekoliko od njih opisuje temeljne tehnike i metode koje se mogu koristiti za olakšavanje tri osnovna koraka implementacije SER-a. S obzirom na to da je tehnologija u neprestanom razvoju te se gotovo svakodnevno pojavljuju nove metode, fokus u ovom radu bit će na najučestalijim i poznatijim metodama. Ostale novije metode uglavnom su neke od kombinacija postojećih uz dodatak drugih metoda iz srodnih grana, principa ili najboljih praksi. Temeljna podjela bit će na tradicionalne i moderne metode prepoznavanja emocija iz snimke glasa.

5.2.1. Tradicionalne metode prepoznavanja emocija

U tradicionalne metode spada spektralna analiza, iako možda nije metoda u samom smislu riječi jer se koristi više kao generator ulaznih podatak za neuronske mreže, koja se više bazira na analiziranju značajki iz spektralne slike zapisa zvuka [10]. Dok se kod modernih metoda za detektiranje emocija u glasovnom zapisu koristi neki od oblika neuronskih mreža. Spektralna analiza je postupak proučavanja spektra signala kako bi se identificirale njegove frekvencijske komponente, što je temeljno u obradi signala, uključujući analizu govora [10]. U kontekstu prepoznavanja emocija iz govora, spektralna analiza omogućuje detaljno razumijevanje kako različite frekvencijske komponente govora odražavaju emocionalno stanje govornika. Proces korištenja spektralne analize može se pojednostaviti u četiri koraka: [10]

1. **Prikupljanje i pretprocesiranje glasovnih zapisa** - Prvi korak uključuje prikupljanje glasovnih zapisa govora te njihovo pretprocesiranje. U obradu je uključeno, ali ne i nužno filtriranje šuma, normalizaciju glasnoće i segmentaciju govora iz tišine ili pozadinskih zvukova.

2. **Transformacija Signala** - Centralni dio spektralne analize je transformacija vremenskih audio signala u frekvencijsku domenu, obično koristeći Fourierovu transformaciju (FFT) ili srodne tehnike kao što su kratkotrajna Fourierova transformacija (STFT) za analizu promjena spektra tijekom vremena. Ovo omogućava vizualizaciju i kvantifikaciju frekvencijskih komponenti sadržanih u signalu.

3. **Ekstrakcija Značajki -** Na temelju transformiranih podataka, ekstrahiraju se relevantne spektralne značajke koje su pokazatelji emocionalnih stanja. Te značajke sadrže mel-frekvencijske kepstralne koeficijente (**MFCC**), spektralnu energiju, spektralni nagib, spektralnu ravnotežu.

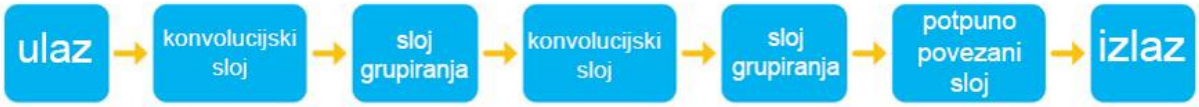
4. **Analiza i Klasifikacija**: Ekstrahirane značajke zatim se koriste za analizu i klasifikaciju emocionalnih stanja. Ovo se može postići koristeći različite metode strojnog učenja, od jednostavnih klasifikatora kao što je k-NN do složenijih modela dubokog učenja. Spektralne značajke služe kao ulazni podaci za ove modele, koji uče prepoznavati obrasce povezane s različitim emocionalnim stanjima.

Osim spektralne analize, spominje se i još jedan pristup - najjednostavnija metoda, odnosno algoritam, a to je K-NN. Algoritam k-najbližih susjeda je nelinearni klasifikator baziran na nadgledanom učenju, koji se oslanja na blizinu kako bi izvršio klasifikaciju ili predviđanje pripadnosti pojedinačne točke podataka nekoj grupi, u našem slučaja emociji iz snimke glasa. K-NN predstavlja jedan od popularnih i najjednostavnijih klasifikatora za klasifikaciju i regresiju koji se danas koriste u strojnom učenju. Iako se algoritam k-NN može koristiti i za probleme regresije i klasifikacije, tipično se upotrebljava kao algoritam za klasifikaciju, polazeći od pretpostavke da se slične točke nalaze blizu jedna drugoj. Za probleme klasifikacije, oznaka klase dodjeljuje se na osnovu glasanja većine - tj. koristi se oznaka koja se najčešće pojavljuje u okolini dane točke podataka. [11]

5.2.2. Moderne metode prepoznavanja emocija

Moderne metode uključuju korištenje neuronskih mreža, neke od najpoznatijih metoda su konvolucijska neuronska mreža (CNN), rekurentna neuronska mreža (RNN), te duboka neuronska mreža (DNN). Osim neuronskih mreža postoje razni algoritmi koji se koriste u modernije vrijeme kao što je i potporni vektorski strojevi (SVM). Ostali algoritmi su manje zastupljeniji, dok s treće strane imamo kombinacije navedenih neuronskih mreža kao DCNN, što bi označavalo duboku konvolucijsku neuronsku mrežu. U nastavku slijede opisi navedenih metoda, tj. klasifikacija, kao i njihove prednosti i nedostatci.

Konvolucijska neuronska mreža predstavlja kategoriju modela strojnog učenja, odnosno vrstu algoritma dubokog učenja koji je posebno prikladan za analizu vizualnih podataka. CNN-ovi, ponekad nazivani konvneti, koriste principe linearne algebre, posebice operacije konvolucije, za izdvajanje značajki i prepoznavanje uzoraka unutar slika. Iako se CNN-ovi pretežno koriste za obradu slika, moguće ih je prilagoditi i za rad s audio podacima te drugim vrstama signala. Arhitektura CNN-a nadahnuta je obrascima povezivanja ljudskog mozga - posebno vizualnim korteksom, koji igra ključnu ulogu u percepciji i obradi vizualnih podražaja. Za razliku od starijih oblika neuronskih mreža, koje su često morale obraditi vizualne podatke na fragmentiran način, koristeći segmentirane ili slike niže rezolucije, sveobuhvatan pristup CNN-a prepoznavanju slika omogućava mu nadmašivanje tradicionalnih neuronskih mreža u nizu zadataka povezanih sa slikama, a u manjoj mjeri, i u obradi govora i audio podataka. CNN-ovi koriste niz slojeva, od kojih svaki detektira različite značajke ulazne slike. Ovisno o složenosti svoje namjene, CNN može sadržavati desetke, stotine ili čak tisuće slojeva, pri čemu svaki gradi na rezultatima prethodnih slojeva kako bi prepoznao detaljne uzorke. Proces počinje odabirom filtera, dizajniranog za detekciju određenih značajki, preko ulazne slike ili audio zapisa, proces poznat kao operacija konvolucije (odatle i dolazi naziv "konvolucijska neuronska mreža"). Rezultat ovog procesa je mapa značajki. Dobivenu mapu zatim prosljeđujemo kao ulaz za sljedeći sloj, omogućavajući CNN-u postupno izgradnju hijerarhijske reprezentacije glasovnog zapisa. [12]



Slika 8. Prikaz strukture CNN-a. [12]

Rekurentna neuronska mreža vrsta je umjetne neuronske mreže koja se koristi za obradu sekvencijalnih podataka ili podataka vremenskih serija. Ovi algoritmi dubokog učenja

često se primjenjuju na temporalne probleme, poput prijevoda jezika, obrade prirodnog jezika, prepoznavanja govora i opisivanja slika. Poput CNN-a, rekurentne neuronske mreže koriste podatke za učenje, a odlikuju se svojom "memorijom" jer koriste informacije iz prethodnih ulaza kako bi utjecale na trenutni ulaz i izlaz. Za razliku od tradicionalnih dubokih neuronskih mreža koje pretpostavljaju da su ulazi i izlazi neovisni jedni o drugima, izlaz rekurentnih neuronskih mreža ovisi o prethodnim elementima unutar sekvence. Iako bi budući događaji također bili korisni u određivanju izlaza dane sekvence, jednosmjerne rekurentne neuronske mreže ne mogu uzeti u obzir te događaje u svojim predviđanjima. Još jedna karakteristična značajka rekurentnih mreža je da dijele parametre kroz svaki sloj mreže. Unatoč tome, te se težine i dalje prilagođavaju kroz procese propagacije unatrag i gradijentnog spusta kako bi se olakšalo pojačano učenje. [13]

Duboko učenje podskup je strojnog učenja koje koristi višeslojne neuronske mreže, nazvane duboke neuronske mreže, za simuliranje složene sposobnosti donošenja odluka ljudskog mozga. Većina umjetne inteligencije (AI) koja danas upravlja našim životima temelji se na nekoj formi dubokog učenja. Prema strogoj definiciji, duboka neuronska mreža je neuronska mreža s tri ili više slojeva. U praksi, većina DNN-ova ima mnogo više slojeva. DNN-ovi se treniraju na velikim količinama podataka kako bi identificirali i klasificirali pojave, prepoznali uzorke i odnose, procijenili mogućnosti ishoda te napravili predviđanja i odluke. Dakle vrijedi pravilo više je bolje - neuronska mreža s jednim slojem može napraviti korisna, aproksimativna predviđanja i donositi odluke, ali s dodatnim slojevima u DNN-u te odluke i ishodi imat će veću točnost i bit će precizniji. Neke od primjena DNN-a su aplikacije i usluge koje omogućuju automatizaciju, izvodeći analitičke i fizičke zadatke bez ljudske intervencije, detekcije prijevara s kreditnim karticama - kao i tehnologije koje se još uvijek razvijaju, poput autonomnih vozila i generativne AI. [14]

Potporni vektorski stroj je algoritam strojnog učenja pod nadzorom, koji klasificira podatke pronalaskom optimalne linije ili hiperplohe koja maksimizira udaljenost između svake klase u N-dimenzionalnom prostoru. SVM razvijeni su 1990-ih godina od strane Vladimira N. Vapnika i njegovih kolega, a svoj su rad objavili u članku pod nazivom "Metoda potpornih vektora za aproksimaciju funkcija, procjenu regresije i obradu signala" 1995. godine. SVM se često koristi unutar problema klasifikacije. Funkcioniraju tako da razlikuju dvije klase pronalaskom optimalne hiperplohe koja maksimizira margine između najbližih točaka podataka suprotnih klasa. Broj značajki u ulaznim podacima određuje je li hiperploha linija u 2-D prostoru ili ravnina u n-dimenzionalnom prostoru. Budući da se može pronaći više hiperploha koje razlikuju klase, maksimiziranje margine između točaka omogućava algoritmu pronalaženje najbolje granice odlučivanja između klasa. To zauzvrat omogućava dobru generalizaciju na

nove podatke i točne predikcije klasifikacije. Linije koje su susjedne optimalnoj hiperplohi poznate su kao potporni vektori jer ti vektori prolaze kroz točke podataka koje određuju maksimalnu marginu. Algoritam SVM na široko se koristi u strojnom učenju jer može rješavati i linearne i nelinearne zadatke klasifikacije. Međutim, kada podaci nisu linearno odvojivi, koriste se kernel funkcije za transformaciju podataka u višedimenzionalni prostor kako bi se omogućila linearna separacija. [15]

Sada kada znamo više o pojedinoj metodi ukratko ćemo sumirati i njihove prednosti i nedostatke te prikazati unutar tablice.

	Prednosti	Nedostatci
DNN	(+) DNN-ovi su fleksibilni i mogu se koristiti za širok spektar zadataka, uključujući klasifikaciju, regresiju i više. (+) Mogu automatski naučiti relevantne značajke iz sirovih podataka, smanjujući potrebu za ručnim inženjeringom značajki.	(-) Zahtijevaju velike količine označenih podataka za učinkovito treniranje. (-) Kao i većina dubokih učenja modela, mogu biti teški za interpretaciju i razumijevanje kako donose odluke.
CNN	(+) Izuzetno su dobri u radu s vizualnim podacima, kao što su slike i video, zbog svoje sposobnosti prepoznavanja uzoraka i struktura. (+) Zahvaljujući konvolucijskim operacijama, CNN-ovi mogu efikasno smanjiti dimenzionalnost slika čuvajući pritom bitne informacije.	(-) Iako su vrlo moćni za vizualne zadatke, nisu jednako fleksibilni kao DNN-ovi za ne-vizualne zadatke. (-) Mogu biti složeniji za implementaciju i treniranje od nekih drugih tipova neuronskih mreža.
RNN	(+) Mogućnost RNN-ova da "pamte" prethodne ulaze omogućava im da razumiju kontekst u sekvencijalnim podacima. (+) Izvrsni su za rad s sekvencijalnim podacima poput vremenskih serija,	(-) Tradicionalni RNN-ovi mogu imati problema s učenjem dugih ovisnosti u podacima zbog nestajućih ili eksplorirajućih gradijenata

	govora ili teksta zbog svoje sposobnosti da zadrže informacije iz prethodnih koraka.	(-) RNN-ovi mogu biti teži za treniranje i optimizaciju zbog svoje sekvencijalne prirode i problema s gradijentima.
K-NN	(+) k-NN je intuitivno jednostavan algoritam, s malom ili nikakvom potrebom za treniranjem modela u tradicionalnom smislu. (+) Može se koristiti za klasifikaciju, regresiju i pretraživanje sličnosti, što ga čini prilagodljivim raznim zadacima. (+) Za manje skupove podataka, k-NN može biti vrlo efikasan i točan.	(-) k-NN može postati izuzetno spor s porastom veličine skupova podataka zbog potrebe za izračunom udaljenosti između testnog uzorka i svih uzoraka u treniranom skupu. (-) Performanse mogu biti značajno negativnije, ako skup podataka sadrži značajke koje nisu relevantne za zadatak. (-) Pronalaženje optimalnog broja 'k', broj susjeda može biti teško i zahtijeva pažljivo testiranje i validaciju.
SVM	(+) SVM je posebno robustan za visokodimenzionalne podatke, što ga čini idealnim za složene zadatke klasifikacije. (+) Uz pravilan odabir kernela, SVM može efikasno rješavati ne-linearne probleme, pružajući precizne granice odlučivanja. (+) Sposobnost minimizacije greške generalizacije, posebno kroz koncept maksimalne margine, čini SVM vrlo otpornim na prenaučenost u usporedbi s nekim drugim modelima.	(-) Odabir pravog kernela i podešavanje njegovih parametara može biti izazovno i zahtijeva eksperimentiranje. (-) Treniranje SVM modela može biti računski zahtjevno za veoma velike skupove podataka. (-) Modeli SVM mogu biti teški za interpretaciju, posebno u usporedbi s nekim jednostavnijim modelima.

Tablica 1. Prikaz prednosti i nedostataka pojedine metode. [10]-[15]

Iako svaki tip neuronske mreže ima svoje specifične prednosti i nedostatke, izbor između DNN, CNN, RNN, K-NN i SVM često ovisi o specifičnom zadatku, vrsti podataka s kojima se radi i specifičnim ciljevima projekta. U sljedećem poglavlju analizirat će se upravo njihova učinkovitost te vidjeti koja od tih metoda je optimalna za određivanje emocija iz snimke glasa.

5.3. Pregled znanstvenih članaka na temu i usporedba uspješnosti metoda

U ovom poglavlju analizirat će se znanstveni radovi te njihovi rezultati. Među mnoštvom izvrsnih i opširnih radova izabrani su oni koji su imali najviše dodirnih točaka i poveznica s temom ovog diplomskog rada.

Analizu ćemo započeti s pregledom znanstvenog članka koji daje uvid u sve moguće pristupe strojnog učenja za prepoznavanje emocija. Autori [16] objašnjavaju koje sve vrste baza podataka su koristili u njihovoj analizi.

Baza podataka	#em	jezik	AV	Nat.	#ms	#fs	#ut	#sent
DES [36,37]	5	Danski	n	glumci	2	2	13	
EMODB [18]	7	Njemački	n	glumci	5	5	10	800
eNTERFACE [89]	6	Engleski	y	prirodno	34	8	-	-
IEMOCAP [19]	10	Engleski	y	glumci	5	5	-	-
SAVEE [47]	7	Engleski	y	glumci	4	-	120	480
Thai DB [128]	6	Tai	y	glumci	3	3	972	5832
INTER1SP [87]		Španjolski	n	glumci	1	1	184	6040
TESS [34]	7	Engleski	n	glumci	-	2	200	2800
RAVDESS [79]	8	Engleski	y	glumci	12	12	2	1440
JL-Corpus [57]	10	Engleski	n	glumci	4	0	-	-
MSP-PODCAST [80]	8	Engleski	n	prirodno	-	-	-	18000

Slika 9. Korištene baze podataka [16]

Zatim slijedi popis u obliku tablice gdje zaglavlje čine klasifikatori, a stupac metode za ekstrakciju značajki, a njihov presjek označava referencu na literaturu gdje se koriste.

Ekstraktori značajki	Klasifikatori										
	LSVM	VSVM	RSVM	TSVM	LR	MLP	ELMDT	BN	GMM	EDT	kNN
MFCC	[40,93,88]	[63,108,46,133,23,27,123,24, 41,116,2,161,90,161,160,109]	[98,93]	[150]	[63,155,167]	[24,54,109,55,22,28]			[97,158,160,127,66]	[41,112,54]	[161,25,2,160,16,111,140]
delta MFCC		[132]			[167]						
LPCC		[132,116]	[100]							[31]	
energy	[40]	[27,108,116]				[24]					
LSP	[40]										
ZCR	[40]	[41,90,116]								[41]	
pitch		[27,108]				[28]			[97]		
arbitrary/ optimized features		[3,149,23,64,10,114,148,95,137,110]	[165]	[150]	[72]	[3,22,165]	[62]	[3]	[11,115,148,82]	[165,60,148]	[165,81]
LDA		[24,78,85]	[98]			[78,24]	[78]				
PCA		[24,10]	[98]			[24]					
HuM		[132]									
Wavelet	[144]	[2,118]	[100,144]						[66]		[2]
NNMF					[125]						[51]

Slika 10. Ovisnosti klasifikatora i ekstraktora značajki. [16]

Zaključno autori govore o problemu dostupnosti javnih baza podataka koje su se koristile za treniranje modela. Iako postoje nešto veće baze i dalje nisu dorasle zadatku. Autori pozivaju na zajedničko dijeljene podataka i testiranje više baza podataka kako bi se riješio problem nedostataka podataka. Na istu temu i do sličnih zaključaka dolaze i drugi autori [17], naglašavajući još neke od problema kao što su manjak govornika s raznih govornih područja i jezika, korištenje isključivo pojedinačnih modela, a ne kombinacije modela te ne postojanju trenda za prepoznavanje emocija na pojedinim govornim područjima i slično.

Sljedeće na redu su znanstveni radovi temeljeni na nekoj od varijanti CNN-a. U prvom radu [9] autori koriste python programski jezik uz uporabu čistog CNN-a. U svome radu primjenjuju parametre od Murray and Arnott 1993 [9].

	Ljutnja	**Sreća**	**Tuga**	**Strah**	**Gađenje**
Brzina	Nešto brže	Brže ili sporije	Nešto sporije	Mnogo brže	Vrlo mnogo brže
Prosječna visina glasa	Vrlo mnogo viša	Mnogo viša	Nešto niža	Vrlo mnogo viša	Vrlo mnogo niža
Raspon glasa	Mnogo širi	Mnogo širi	Nešto uži	Mnogo širi	Nešto širi
Intenzitet	Viši	Viši	Niži	Normalan	Niži
Kvaliteta glasa	Dahav, prsni	Dahav, zvučan ton	Rezonantan	Nepravilan glas	Mrmljavi prsni ton
Promjene glasa	Nagle na naglašenim	Glato, uzlazne promjene	Silazne promjene	Normalne	Široke, silazne završne promjene
Artikulacija	Napeta	Normalna	Mumljanje	Precizno	Normalna

Slika 11. Parametri emocija i govora. [9]

Evaluaciju su radili na uzorku od 30 glasovnih zapisa, a rezultati su sljedeći:

Sretno	Strah	Ljutnja	Tužno	Gađenje	Iznenađeno
71	75	68	74	67	69

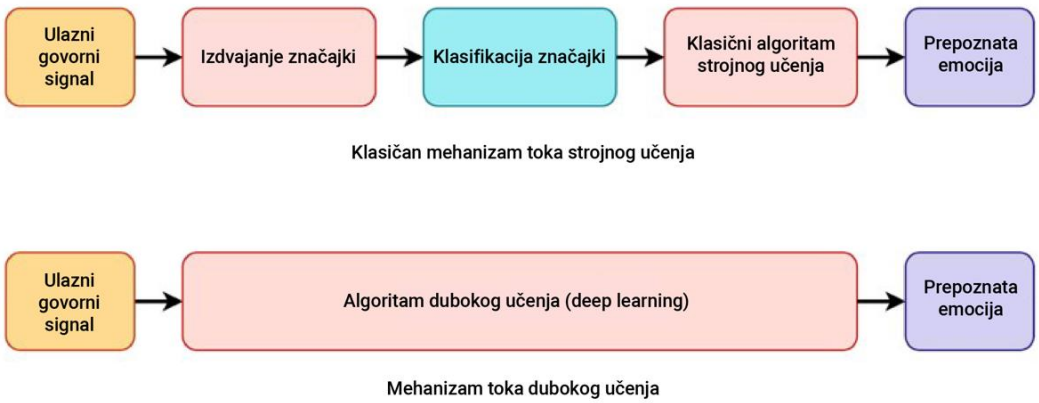
Slika 12. Rezultati CNN-a. [9]

Model je postigao medijan točnosti od 71.37 % koji je relativan i usporediv s njihovim izvorima. Također su usporedili i svoj model s ostalim metodama kao što je SVM i KNN koji su imali nešto bolje rezultate za emocije tuge i ljutnje. Kod autora [18] vidljivo je spajanje CNN modela s LSTM modelom. U tom radu, podaci se normaliziraju koristeći normalizaciju srednjeg kvadratnog korijena, zatim se računa i klasificira broj datoteka za svaku emociju. Prema konvenciji klasifikacije sentimenta, skup za treniranje, testiranje i validaciju čine ukupno 64 %, 20 % i 16 % od 535 podataka. Potrebno je standardizirati podatke za CNN-LSTM model jer se dužina pretvorenih podataka razlikuje, gdje je najduži 143652, a najkraći 19608, te je potrebno popuniti nedostajuće podatke. Normalizirani audio podaci segmentirani su na fiksnu duljinu od 16000 Hz i dopunjeni nulama. Naposljetku, svaki audio može formirati dvodimenzionalne podatke kao ulaz za CNN+LSTM model. Autori se prvenstveno fokusiraju na četiri glavne emocije - ljutnje, sreće, tuge i neutralnosti. Ostale emocije u skupu podataka uklonjene su zbog nedostatka dovoljno podataka koji mogu ometati performanse korištenih modela. CNN-LSTM model korišten za prepoznavanje emocija sastoji se od četiri sloja CNN-a, jednog sloja LSTM-a i potpuno povezanih slojeva. Model pokazuje varijabilnu točnost na temelju podjele skupa za testiranje i treniranje EMO-DB skupa podataka, s točnošću prepoznavanja emocija od 88,50 % kada se koriste četiri emocije za treniranje i predviđanje s podjelom 0,2 za testiranje. Dakle znatno bolje od samog CNN modela kod autora [9]. Treći i zadnji znanstveni rad koji ulazi u analizu na području CNN, kombinira CNN s „multi-head“ konvolucijskim transformatorom [19]. Dobiveni rezultati su vrlo približni modelu CNN+LSTM, ovisno o tome na kojoj bazi podataka se vršilo testiranje.

			RAVDESS			IEMOCAP		
Ref#	**Usporedbe**	**Ulazne Značajke**	**Točnost**	**Preciznost**	**F1 Rezultat**	**Točnost**	**Preciznost**	**F1 Rezultat**
[56]	BE-SVM	Spektralne Značajke	75.69	74.00	73.34	-	-	-
[85]	GResNets	Spektralne Značajke	64.48	65.32	63.11	-	-	-
[86]	MLT-DNet	Prostorne Značajke	-	-	-	73.01	74.00	73.00
[57]	Deep-BLSTM	Prostorne + Temporalne	77.02	76.00	77.00	72.50	73.00	72.00
[74]	1D-CNN	Spektralne Značajke	71.61	-	-	64.30	-	-
[66]	DS-CNN	Prostorne Značajke	79.50	81.00	84.00	78.75	86.00	82.00
[60]	DeepNet	Prostorne + Temporalne	-	-	-	77.00	76.00	76.00
[68]	Att-Net	Prostorne Značajke	80.00	81.00	80.00	78.00	78.00	78.00
Naš	CTENet	Prostorne + Temporalne	82.31	81.75	84.37	79.42	74.80	82.20

Slika 13. Usporedba CNN modela s transformatorom s ostalim modelima. [19]

Nakon CNN modela, koje možemo svrstati kao potkategoriju DNN-a, dolaze nam ostale vrste DNN modela. Autori, u svom radu [20] vrlo detaljno prikazuju način na koji funkcioniraju DNN-ovi te njihovu arhitekturu, a najbolju razliku između DNN-a i ostalih načina strojnog učenja možemo vidjeti na slici ispod.



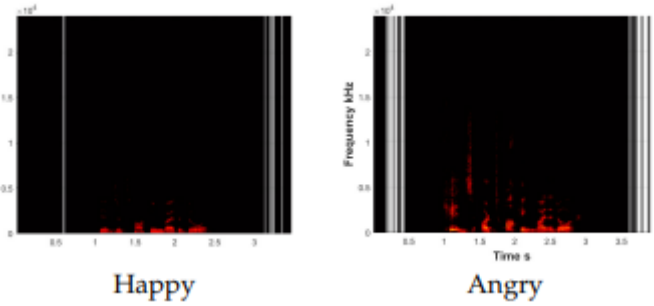
Slika 14. Usporedba DNN mehanizma s ostalim mehanizmima strojnog učenja. [20]

Proširivanjem slojeva CNN-a kao duboke neuronske mreže, autori su postigli zapanjujući rezultat od gotovo 100 % preciznosti u određivanju emocija. Jedini nedostatak ove metode je taj što model ne može određivati s tako velikom preciznosti ostale emocije koje nisu navedene na slici ispod.

Algoritmi	Ljutnja	Sreća	Tuga
k-najbliži susjed	93%	55%	77%
Linearna diskriminantna analiza	68%	49%	72%
Stroj potpornih vektora	74%	70%	93%
Regularizirana diskriminantna analiza	83%	73%	97%
Duboka konvolucijska neuronska mreža	99%	99%	96%

Slika 15. Komparacija modela DNN-a u određivanju osnovnih emocija. [20]

Za kraj preostaje spomenuti i rad koji se temelji na tradicionalnim metodama spektralne analize, odnosno audio matching-a [21]. Za očekivati je da će imati znatno slabije rezultate naspram modernijih tehnika, no nije ništa manje bitan jer je upravo on bio preteča istraživanja današnjih neuronskih mreža. Uz to autori uspoređuju audio matching s ostalim modelima baziranim na neuronskim mrežama, te dolaze do zaključka kako je LTSM model superiorniji naspram RNN modela, jer ne traži dodatna stanja memorije.



Slika 16. Spektralni prikaz izgovaranja iste rečenice sa srećom (lijevo) i s ljutnjom (desno) [21]

Na slici iznad prikazana su dva spektra, na desnom se može primijetiti više glasovnih karakteristika kod izgovaranja rečenice ljutitim tonom, za razliku od izgovaranja sretnim tonom/ glasom. Sumirani rezultati ovog znanstvenog članka prikazani su sljedećom slikom.

Baza podataka	Klasifikatori	Sreća	Ljutnja	Neutralno	Ukupno
Emotion	Naive Bayes	0.35	0.49	0.41	0.42
	GMM	0.46	0.49	0.39	0.45
	Wav2vec	0.54	0.5	0.28	0.44
	Random Forest	0.57	0.63	0.38	0.56
	k-NN	0.59	0.6	0.52	0.58
	LSTM	0.52	0.67	0.48	0.56
	SAM	0.76	0.81	0.77	**0.78**
YouTube	Naive Bayes	0.31	0.63	0.01	0.3
	GMM	0.51	0.42	0.27	0.40
	Wav2vec	0.61	0.27	0	0.29
	Random Forest	0.39	0.6	0.01	0.32
	k-NN	0.41	0.52	0.12	0.34
	LSTM	0.4	0.33	0.15	0.29
	SAM	0.68	0.54	0.1	**0.44**

Slika 17. Rezultati i usporedbe tradicionalnih metoda s modernima. [21]

Zanimljivo je da u većini istraživanja uspješnost prepoznavanja emocija iznosi 70-80 %, uz iznimku DCNN koji je skoro sto postotan u određivanju emocija. Dok je kod ljudi postotak točnosti u prepoznavanju emocija nešto niži od prosjeka metoda strojnog učenja – 60 % [9]. Što dovodi do pitanja znaju li strojevi bolje prepoznati naše emocije od nas samih?

6. Implementacija prepoznavanja emocija iz glasa

U praktičnom dijelu diplomskog rada proći će se koraci od instalacije programa i biblioteka potrebnih za prepoznavanje emocija iz snimke glasa do objašnjenja kako stvari funkcioniraju te usporediti dvije implementirane metode za prepoznavanje emocija – k-najbližih susjeda te konvolucijske neuronske mreže.

Za izradu programa korišteno je sljedeće okruženje, alati i biblioteke:

- Windows 10
- Visual studio code
- Python 3.8.0
- Python pip verzija 24.0
- librosa
- tensorflow
- matplotlab
- seaborn
- keras
- scikit-learn
- RAVDESS baza podataka zvučnih zapisa

Prvi korak je instalacija python okruženja. Nakon toga u terminalu pomoću pip instaliramo potrebne biblioteke:

```
pip install librosa tensorflow keras scikit-learn
```

Pričekamo da se paketi instaliraju na računalo. Moguće da neka od verzija odstupa, međutim pip bi trebao sve automatski posložiti. Librosa se koristi za ekstrakciju značajki zvuka, vidjet ćemo kasnije na primjeru kako je to postignuto pomoću već spomenute metode MFCC. Tensorflow kao i Keras su biblioteke za kreiranje i rad s neuronskim mrežama, dok se scikit-learn koristi za treniranje modela i kodiranje podataka zvuka.

6.1. Ekstrakcija značajki

Prilikom ekstrakcije značajki koristi se značajka biblioteke Librosa – feature.mfcc. Pomoću nje prvo učitamo zvučni zapis, uzimajući podatke te sample rate, koji se dalje koristi za ekstrakciju značajki zvuka. Mel-frekvencijske kepstralne koeficijente (MFCC) smo morali zbog treniranja neuronskih mreža svesti na istu dužinu, a to se radi u vremenskim koracima i

to predstavlja broj 180 (slika 19), čemu je otprilike ekvivalentno 4 sekunde, ovisno o sample rate-u zapisa zvuka. Oni uzorci zvučnih zapisa koji su duži od toga bit će odrezani do dužine od 180 vremenskih koraka. Metoda K-NN nema potrebu za optimizacijom jer radi na drugom principu.

```
def load_data(dataset_path):
    features, emotions = [], []
    for file in os.listdir(dataset_path):
        emotion_code = int(file.split("-")[2])
        emotion = {1: 'neutral', 2: 'calm', 3: 'happy', 4: 'sad', 5:
'angry', 6: 'fearful', 7: 'disgust', 8: 'surprised'}.get(emotion_code,
None)
        if emotion is not None:
            path = os.path.join(dataset_path, file)
            data, sampling_rate = librosa.load(path)
            mfccs = np.mean(librosa.feature.mfcc(y=data,
sr=sampling_rate, n_mfcc=40).T, axis=0)
            features.append(mfccs)
            emotions.append(emotion)
    return np.array(features), np.array(emotions)
```

Slika 18. Prikaz ekstrakcije značajki korištenjem MFCC kod K-NN.

```
def load_data(dataset_path):
    features, emotions = [], []
    for file in os.listdir(dataset_path):
        parts = file.split("-")
        if len(parts) >= 3:
            try:
                emotion_code = int(parts[2])
                emotion = {1: 'neutral', 2: 'calm', 3: 'happy', 4:
'sad', 5: 'angry', 6: 'fearful', 7: 'disgust', 8:
'surprised'}.get(emotion_code, None)
                if emotion is not None:
                    path = os.path.join(dataset_path, file)
                    data, sampling_rate = librosa.load(path)
                    mfccs = librosa.feature.mfcc(y=data,
sr=sampling_rate, n_mfcc=40)
                    if mfccs.shape[1] > 180:
                        mfccs = mfccs[:, :180]
                    else:
                        mfccs = np.pad(mfccs, ((0, 0), (0, 180 -
mfccs.shape[1])), "constant")
                    features.append(mfccs)
                    emotions.append(emotion)
            except ValueError:
                print(f"Skipping file with unexpected format: {file}")
    return np.array(features), np.array(emotions)
```

Slika 19. Prikaz ekstrakcije značajki korištenjem MFCC kod CNN.

6.2. Klasifikacija

Kod klasifikacije emocija koristimo 8 emocija koje su došle s RAVDESS bazom podataka. RAVDESS baza podataka klasificira neutralno, opušteno, sretno, tužno, strah, gađenje i iznenađeno emocionalno stanje. Baza se sastoji od 1440 zvučnih zapisa od 24 osobe – pola muškaraca pola žena. Svatko izgovara po dvije rečenice uz dva ponavljanja iste rečenice. Dakle, model u oba slučaja, K-NN i CNN, je treniran upravo na tim podacima. Ostale open source baze su bile znatno manje (EmoDB, SAVEE…) ili nisu imale engleske govornike pa nisu uzete u obzir. Naravno čim je veći broj podataka na kojem se model trenira to će i rezultati biti bolji. Na slikama 20 i 21 možemo vidjeti kako izgleda treniranje i postavke modela.

```
features, emotions = load_data('data')

label_encoder = LabelEncoder()
emotions_encoded = label_encoder.fit_transform(emotions)
emotions_one_hot = to_categorical(emotions_encoded)

features = np.expand_dims(features, axis=-1)

X_train, X_test, y_train, y_test = train_test_split(features, emotions_one_hot, test_size=0.2, random_state=42)

model = Sequential([
    Conv2D(32, kernel_size=(3, 3), activation='relu', input_shape=(40, 180, 1)),
    MaxPooling2D(pool_size=(2, 2)),
    Dropout(0.25),
    Conv2D(64, (3, 3), activation='relu'),
    MaxPooling2D(pool_size=(2, 2)),
    Dropout(0.25),
    Flatten(),
    Dense(128, activation='relu'),
    Dropout(0.5),
    Dense(len(np.unique(emotions_encoded)), activation='softmax')
])

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
model.fit(X_train, y_train, batch_size=32, epochs=100, validation_data=(X_test, y_test))

loss, accuracy = model.evaluate(X_test, y_test)
model.save('cnn.h5')
with open('label_encoder.pkl', 'wb') as file:
    pickle.dump(label_encoder, file)
print(f"Test Loss: {loss}, Test Accuracy: {accuracy}")
```

Slika 20. Kod za treniranje i postavke modela CNN-a.

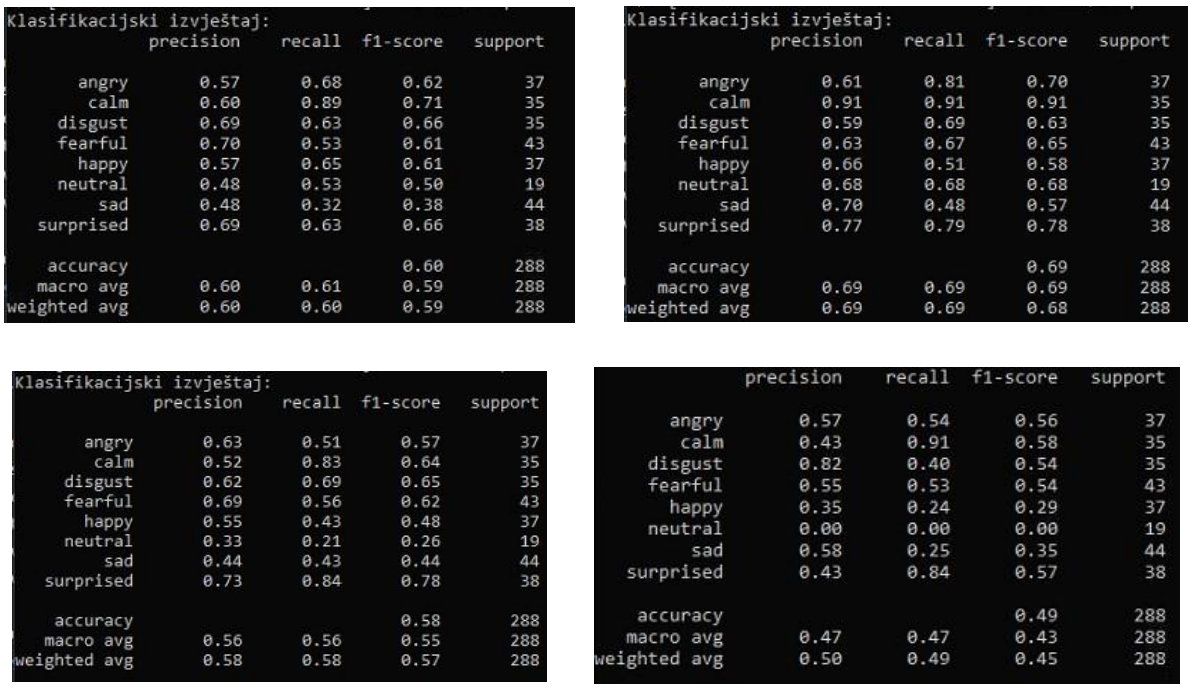
Prilikom treniranja CNN modela bitno je obratiti pažnju na parametre modela. Iz tensorflow biblioteke koristimo sekvencionalnu seriju, jer zvučni zapis spada u tu kategoriju. Zatim odabiremo slojeve koje ćemo koristiti. Na slici 20 koriste se 2D slojevi. Prvi parametar (broj 32) predstavlja broj neurona koji će taj sloj imati, zatim slijedi kernel_size, odnosno filteri. Brojevi (3, 3) predstavljaju veličinu svakog filtera. Parametar input_shape služi za definiranje oblika vučnog zapisa kojeg smo prethodno pripremili pomoću MFCC metode. Activation

parametar ima predefinirane karakteristike – relu i softmax, odabir ovisi o načinu treniranja modela, linearno i ne linearno. Nakon podešavanja prvog sloja dolazi na red optimizacija izlaza tog sloja odnosno metoda MaxPooling2D. Pomoću nje smanjujemo dimenzionalnost i opterećenja prilikom treniranja. Nakon toga dodajemo još jedan 2D sloj, ovoga puta nije potrebno definirati input_shape, ali smo povećali dvostruko broj neurona koji će taj sloj koristiti. Ponovo smanjujemo opseg izlaza te postavljamo Dropout na 0.25. Time smo postigli da model automatski postavlja dio značajki na 0, smanjujući vjerojatnost prenaučenosti. Slijedi korak Flatten() gdje pretvaramo sloj iz 2D u 1D za potpuno povezane slojeve potrebne za Dense(). Na kraju pokrećemo treniranje modela naredbom mode.fit, a batch_size je parametar koji nam govori nakon koliko uzoraka će se ažurirati težine. Napravljeno je nekoliko varijanti testiranja između postavki parametara i korištenja 1D slojeva u startu. Prilikom treniranja modela Neki od rezultata bit će prikazani u nastavku.

```
features, emotions = load_data('data')

label_encoder = LabelEncoder()
emotions_encoded = label_encoder.fit_transform(emotions)

X_train, X_test, y_train, y_test = train_test_split(features, emotions_encoded, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

k = 5
knn = KNeighborsClassifier(n_neighbors=k)

knn.fit(X_train, y_train)

y_pred = knn.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Preciznost k-NN klasifikatora pri k={k}: {accuracy:.2f}")

def extract_features(audio_path):
    data, sampling_rate = librosa.load(audio_path)
    mfccs = np.mean(librosa.feature.mfcc(y=data, sr=sampling_rate, n_mfcc=40).T, axis=0)
    return mfccs

def predict_emotion(audio_path, model, scaler, label_encoder):
    features = extract_features(audio_path)
    features = scaler.transform([features])
    prediction = model.predict(features)
    predicted_emotion = label_encoder.inverse_transform(prediction)
    return predicted_emotion[0]

new_audio_path = 'Snimka.wav'
predicted_emotion = predict_emotion(new_audio_path, knn, scaler, label_encoder)
print(f"Prepoznana emocija iz glasa: {predicted_emotion}")
```

Slika 21. Kod za treniranje i postavke modela K-NN-a.

Kod K-NN metode ovaj proces je nešto jednostavniji. Jedini parametar kojeg je potrebno odrediti je k – klasifikator koji određuje koliko najbližih susjeda se razmatra za odlučivanje o klasifikaciji. Mijenjanjem tog broja možemo dobiti ili izgubiti preciznost, kao i prenaučiti ili premalo naučiti model. Također je vidljivo da se evaluacija i prepoznavanje emocije kod K-NN metode može odmah sprovesti jer se takav model brzo trenira. Dok se treniranje modela kod CNN sprema u posebnu datoteku te naknadno učitavaju kodirane datoteke i istrenirani modeli. Opće postavke modela kao kernel_size, pool_size i batch_size su uzete prema default-u, jer u suprotnom prilikom testiranja (povećavanje i smanjivanje pool-ova) rezultiralo je izrazito niskom preciznosti 15-30%. Potrebno je puno znanja i iskustva u ovom području jer u suprotnom, ako pogriješimo s postavkama možemo naš trenirani model napraviti prenaučenim. U tom slučaju ako predviđamo emocije na treniranom upitu rezultati će biti odlični, no ako pred model stavimo nove rečenice tada će rezultati biti izrazito loši jer model neće moći generalizirati na temelju treniranih podataka na koju se emociju odnosi.



Klasifikacijski izvještaj:

	precision	recall	f1-score	support
angry	0.57	0.68	0.62	37
calm	0.60	0.89	0.71	35
disgust	0.69	0.63	0.66	35
fearful	0.70	0.53	0.61	43
happy	0.57	0.65	0.61	37
neutral	0.48	0.53	0.50	19
sad	0.48	0.32	0.38	44
surprised	0.69	0.63	0.66	38
accuracy			0.60	288
macro avg	0.60	0.61	0.59	288
weighted avg	0.60	0.60	0.59	288

Klasifikacijski izvještaj:

	precision	recall	f1-score	support
angry	0.61	0.81	0.70	37
calm	0.91	0.91	0.91	35
disgust	0.59	0.69	0.63	35
fearful	0.63	0.67	0.65	43
happy	0.66	0.51	0.58	37
neutral	0.68	0.68	0.68	19
sad	0.70	0.48	0.57	44
surprised	0.77	0.79	0.78	38
accuracy			0.69	288
macro avg	0.69	0.69	0.69	288
weighted avg	0.69	0.69	0.68	288

Klasifikacijski izvještaj:

	precision	recall	f1-score	support
angry	0.63	0.51	0.57	37
calm	0.52	0.83	0.64	35
disgust	0.62	0.69	0.65	35
fearful	0.69	0.56	0.62	43
happy	0.55	0.43	0.48	37
neutral	0.33	0.21	0.26	19
sad	0.44	0.43	0.44	44
surprised	0.73	0.84	0.78	38
accuracy			0.58	288
macro avg	0.56	0.56	0.55	288
weighted avg	0.58	0.58	0.57	288

	precision	recall	f1-score	support
angry	0.57	0.54	0.56	37
calm	0.43	0.91	0.58	35
disgust	0.82	0.40	0.54	35
fearful	0.55	0.53	0.54	43
happy	0.35	0.24	0.29	37
neutral	0.00	0.00	0.00	19
sad	0.58	0.25	0.35	44
surprised	0.43	0.84	0.57	38
accuracy			0.49	288
macro avg	0.47	0.47	0.43	288
weighted avg	0.50	0.49	0.45	288

Slika 22. Klasifikacijski izvještaji za CNN ovisno o parametrima.

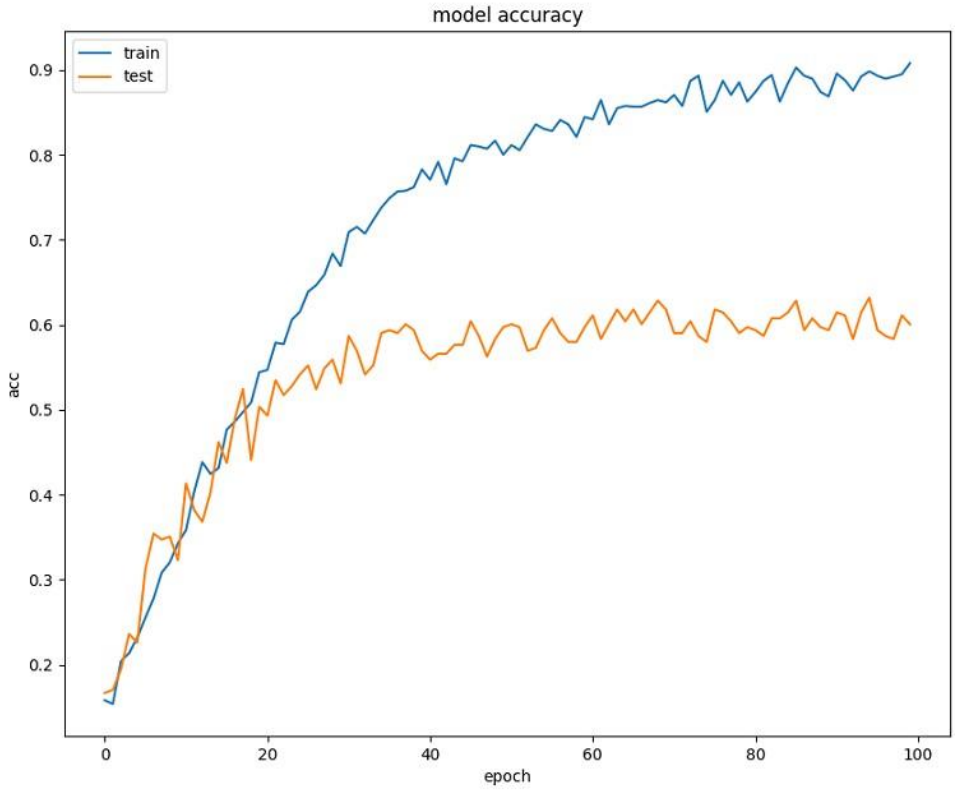
Gore lijevo nalaze se vrijednosti za defaultne postavke, gore desno je izvedba CNN modela koristeći 1D slojeve, dolje lijevo je izvedba s povećanim brojem neurona unutar slojeva, koristeći 2D slojeve. Dok slika dolje desno predstavlja treniranje defaultnog modela kroz 200 epoha. Vidljivo je da najbolje rezultate postiže CNN model koji koristi 1D slojeve, približujući se tako postotku ostalim znanstvenim istraživanjima.

Preciznost k-NN klasifikatora pri k=5: 0.53

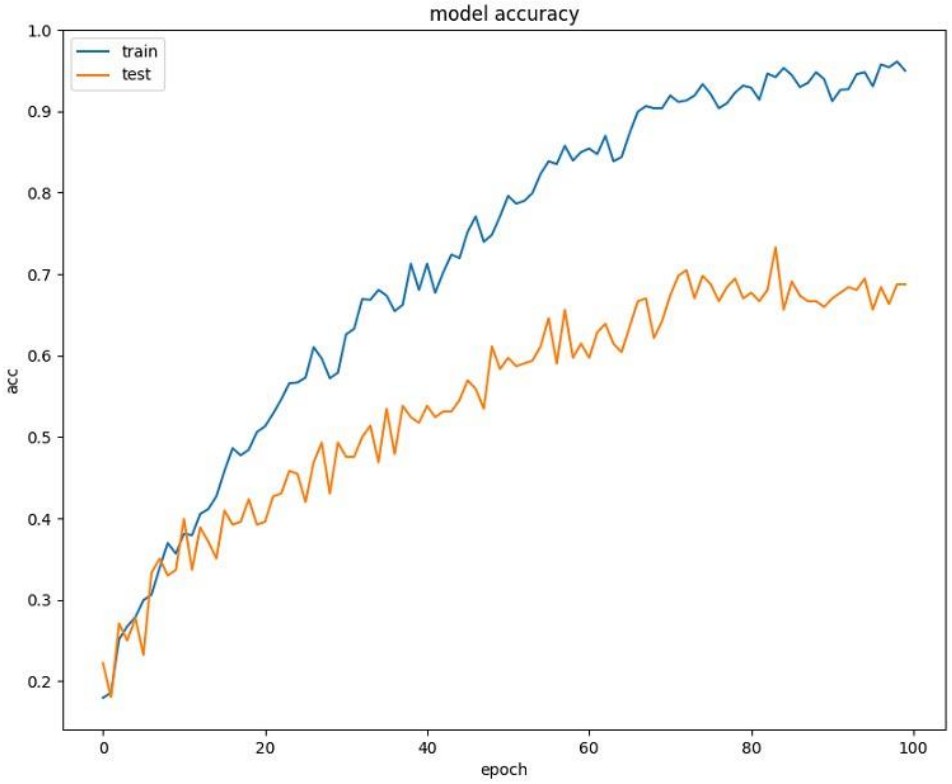
	precision	recall	f1-score	support
angry	0.46	0.62	0.53	37
calm	0.58	0.91	0.71	35
disgust	0.48	0.60	0.53	35
fearful	0.59	0.40	0.47	43
happy	0.43	0.27	0.33	37
neutral	0.53	0.47	0.50	19
sad	0.78	0.41	0.54	44
surprised	0.51	0.63	0.56	38
accuracy			0.53	288
macro avg	0.55	0.54	0.52	288
weighted avg	0.55	0.53	0.52	288

Slika 23. Klasifikacijski izvještaj za metodu K-NN.

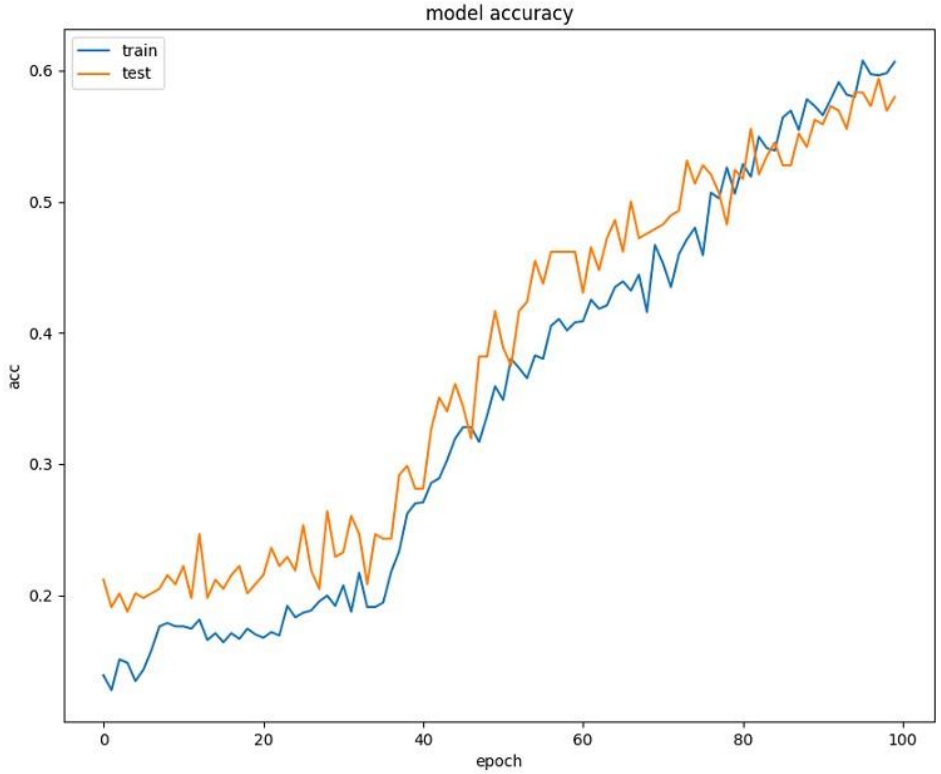
Zanimljivo je još prikazati i grafove procesa treniranja podataka te gubitke. S obzirom na to da K-NN ne radi na istom principu, odnosno nije neuronska mreža ovi grafovi su dostupni samo za CNN model. Pristup podacima omogućuje funkcija iz biblioteke tensorflow. Da bismo iz modela izvukli podatke potrebno je prilikom treniranja model spremiti pod varijablom i nad njim pozvati metodu history.Dobivene podatke zatim prikazujemo pomoću matplotlab biblioteke.



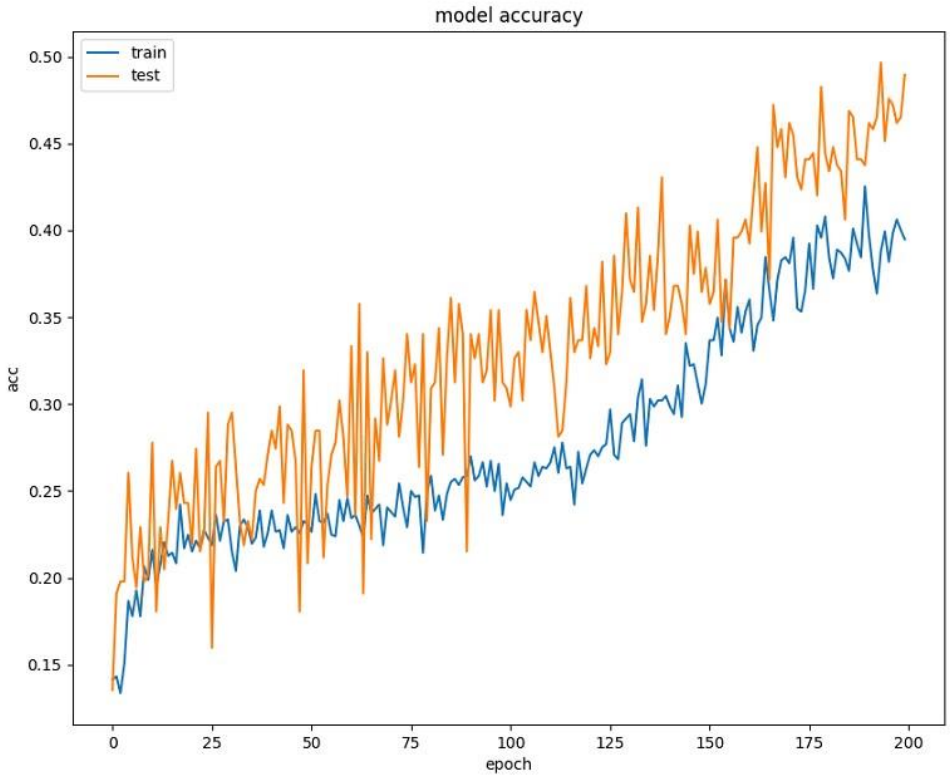
Slika 24. Prikaz preciznosti default modela kroz epohe.



Slika 25. Prikaz preciznosti CNN modela s 1D slojevima.

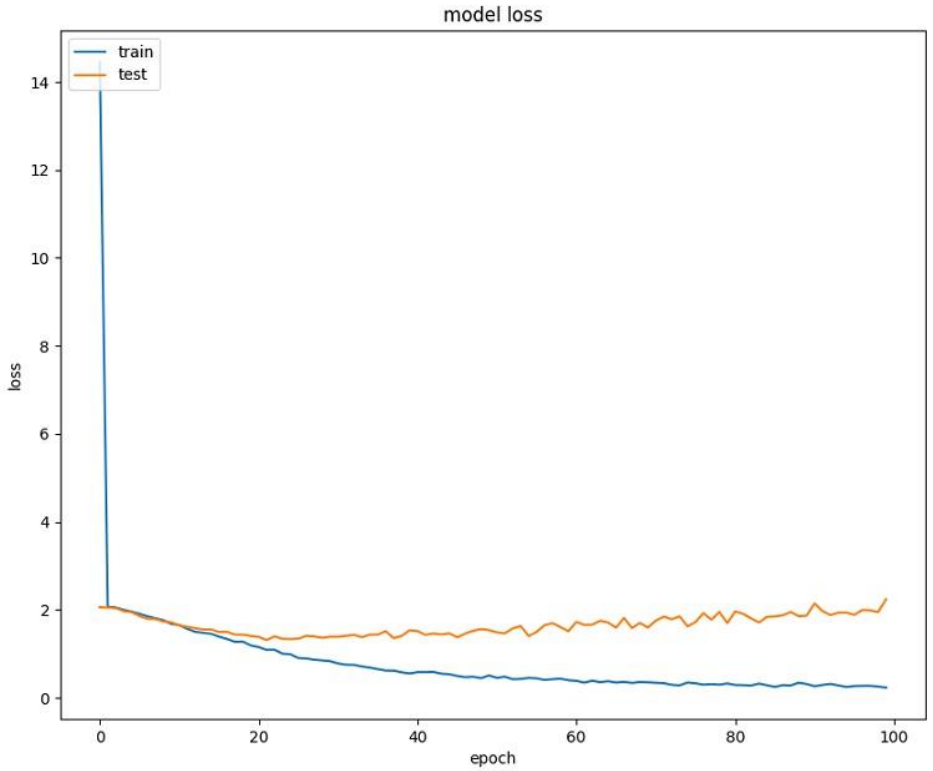


Slika 26. Prikaz preciznosti CNN modela s povećanim neuronima unutar sloja.

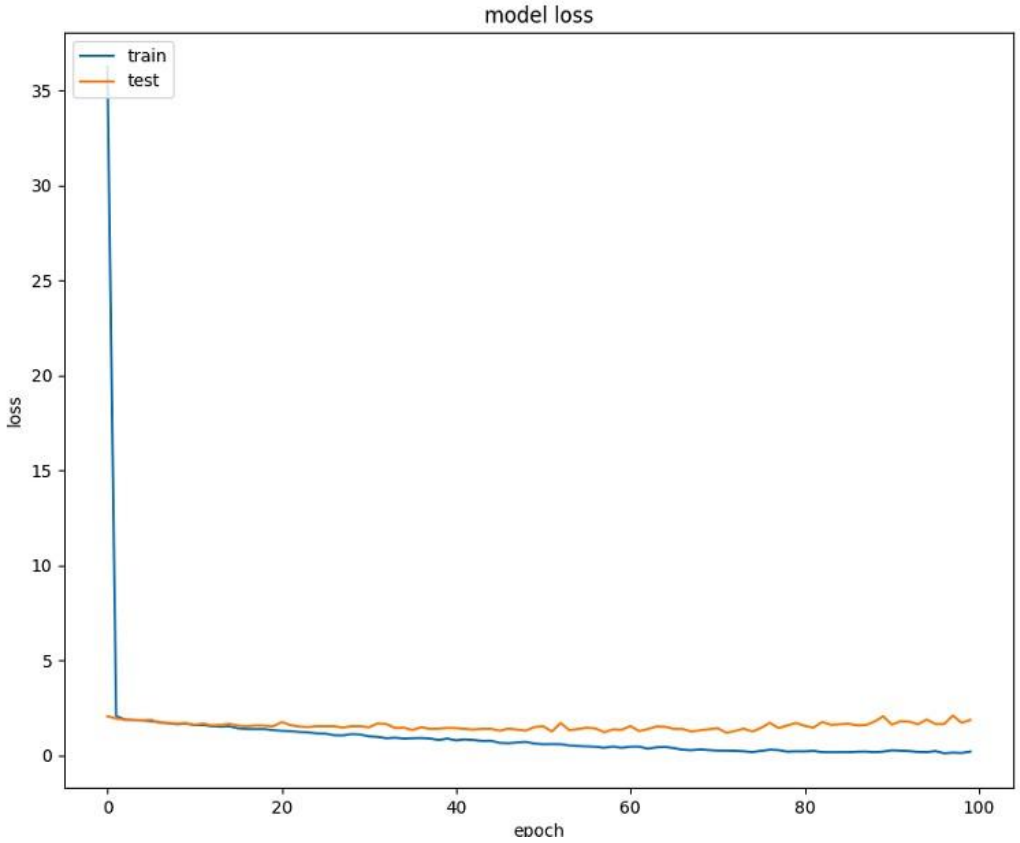


Slika 27. Prikaz CNN default modela kroz 200 epoha treniranja.

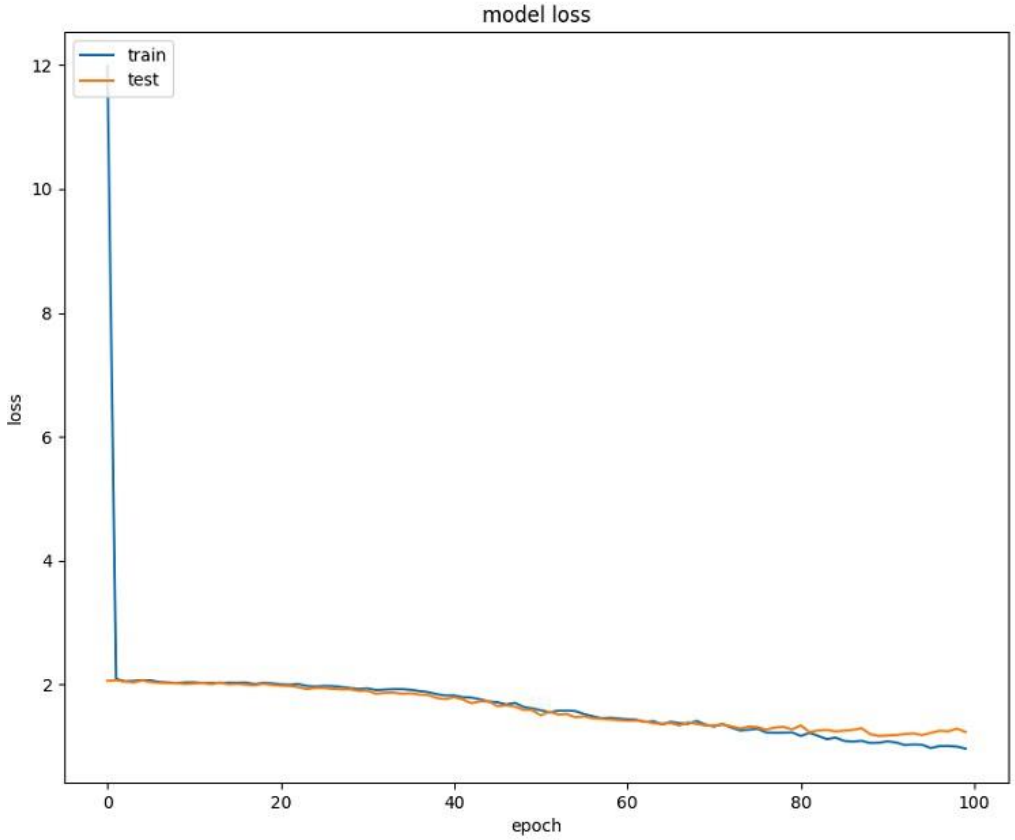
Na prethodnim grafovima jasnije se vidi proces treniranja podataka te kako oscilira razlika između testnih i treniranih podataka. Slično možemo vidjeti i kod evaluacije grešaka.



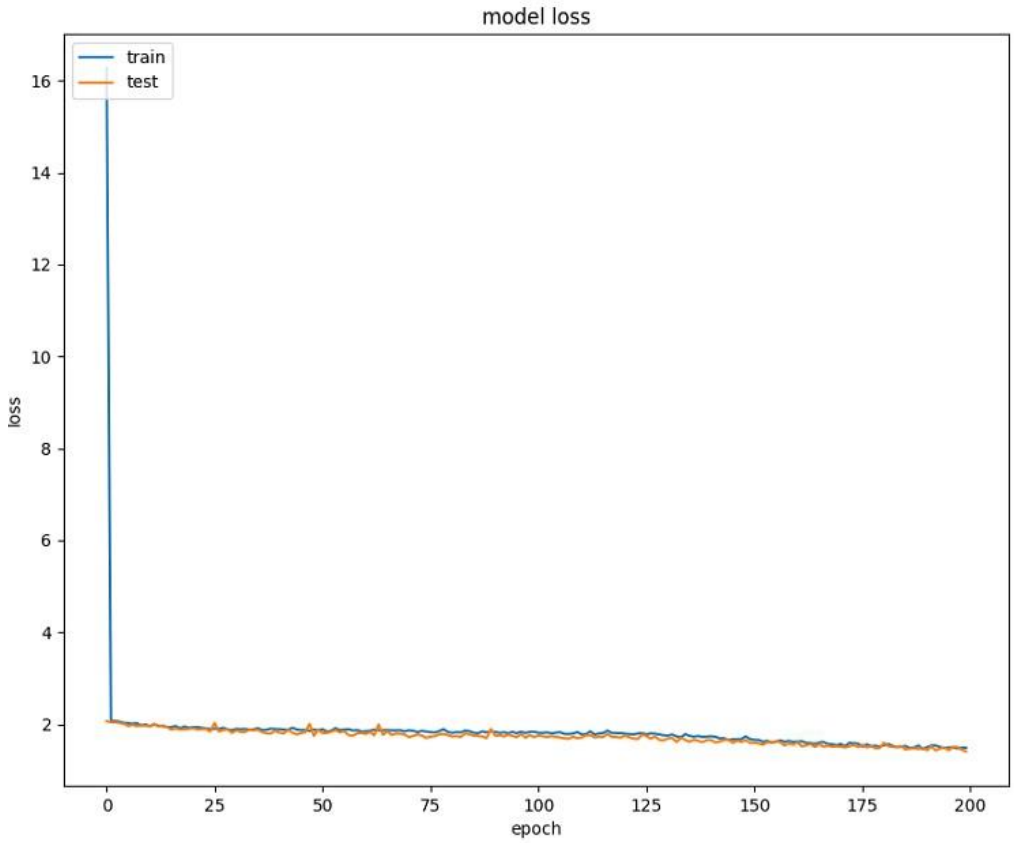
Slika 28. Prikaz gubitka CNN defaultnog modela.



Slika 29. Prikaz gubitka CNN 1D modela.



Slika 30. Prikaz gubitka CNN s povećanim neuronima.



Slika 29. Prikaz gubitka CNN modela kroz 200 epoha.

6.3. Evaluacija

Prilikom treniranja modela na testnom uzorku od 20% ovo su rezultati za pojedini model:

- K-NN – uz k=5 preciznost iznosi 0.53
- CNN 2D - Test Loss: 1.895, Test Accuracy: 0.611
- CNN 1D - Test Loss: 1.937, Test Accuracy: 0.684

Trenirani model za CNN 2D pokazuje odstupanje od 10% naspram istraživanja opisana u prethodnom poglavlju(~70%). Što ne čudi jer je ovo vrlo jednostavan primjer koji je treniran na malom broju podataka i uz uobičajene postavke. Na testiranju modela s osobnim snimkama, od 10 zapisa CNN 2D model je bio precizan u 50 % slučajeva, dok je 30 % proglasio emocijom koje su vrlo slične jedna drugoj po zvučnim karakteristikama, primjerice zvučni zapis je nastojao demonstrirati tugu, no model je proglasio zapis kao gađenje. Preostalih 20 % model je pogrešno procijenio. CNN 1D razlikuje se u tome što je jednu emociju pogodio više od CNN 2D modela te je postotak sljedeći: 60 % - 20 % - 20 %. Što se tiče prepoznavanja emocija

pomoću K-NN algoritma jako veliku ulogu igra odabrani broj k – koji je također bio postavljen na uobičajenu vrijednost 5. K-NN je pokaza slične, ali ipak nešto lošije podatke. Model je bio precizan u 40 % slučajeva, 40 % je prepoznao srodnu emociju, te 20 % potpuno pogriješio. Prema rezultatima, a isto kao i u prethodnim istraživanjima, CNN odnosi pobjedu naspram K-NN. Smatram da su oba modela dobar temelj za daljnju optimizaciju i da sa svojim postotcima mogu konkurirati nekim slabijim metodama. Trenirani model i kod dostupni su na github-u.

```python
import pickle
import librosa
import numpy as np
from tensorflow.keras.models import load_model

def predict_emotion_from_file(file_path, model, label_encoder):
    data, sampling_rate = librosa.load(file_path)
    mfccs = librosa.feature.mfcc(y=data, sr=sampling_rate, n_mfcc=40)
    if mfccs.shape[1] > 180:
        mfccs = mfccs[:, :180]
    else:
        mfccs = np.pad(mfccs, ((0, 0), (0, 180 - mfccs.shape[1])), "constant")
    mfccs = np.expand_dims(mfccs, axis=-1)
    mfccs = np.expand_dims(mfccs, axis=0)
    prediction = model.predict(mfccs)
    predicted_index = np.argmax(prediction, axis=1)
    predicted_emotion = label_encoder.inverse_transform(predicted_index)
    return predicted_emotion[0]

model = load_model('cnn.h5')

with open('label_encoder.pkl', 'rb') as file:
    label_encoder = pickle.load(file)

predicted_emotion = predict_emotion_from_file('Snimka (2).wav', model, label_encoder)
print(f"Predviđena emocija: {predicted_emotion}")
```

Slika 30. Prikaz koda za učitavanje treniranog modela i predviđanje emocija.

Na slici 30 isječak je koda na kojem učitavamo trenirani CNN model te pokušavamo vidjeti hoće li ispravno predvidjeti emocije iz zvučnog zapisa 'Snimka (2).wav'. Snimka (2) je osobni zvučni zapis izgovaranja opuštenim tonom – „tomorrow will be sunny“.

7. Zaključak

U današnjem svijetu neprestanog razvoja tehnologije i dijeljenja informacija, od jednostavnih svakodnevnih odluka do složenih poslovnih strategija, ovaj diplomski rad dao je uvid u potencijale biometrije, s naglaskom na biometriju glasa i prepoznavanje emocija iz snimaka glasa. Kroz rad, otkrili smo kako su biometrijski sustavi, kad su pravilno iskorišteni, moćan alat koji može dramatično olakšati i poboljšati naše živote, smanjiti radno vrijeme i povećati sigurnost u digitalnom dobu. Imajući biometriju glasa kao središte istraživanja, pokazali smo da tehnologija nije samo alat, već prozor u budućnost gdje je granica između čovjeka i stroja sve manje vidljiva. Navedeni su primjeri iz prakse, od jednostavnih aplikacija do složenih biometrijskih sustava, kao i uvidi u karakteristike biometrijskih sustava, njezinu važnost i dublje razumijevanje razlike između verifikacije i identifikacije. Osim toga napravljena je i detaljna analiza metoda koje se mogu koristiti za prepoznavanje emocija iz snimke glasa, gdje daleko prednjači metoda DCNN s jako velikom vjerojatnošću za prepoznavanje svih osnovnih emocija.

Prema tome, biometrijske tehnologije donose velike i ključne značajke u poboljšanju i osiguranju kako osobnih tako i poslovnih podataka. Imajući na umu da ne postoji najbolja biometrija, prošli smo kroz izazove i ograničenja biometrije glasa - od pitanja privatnosti i sigurnosti do tehničkih prepreka koje još uvijek treba prevladati. Ovaj diplomski rad nije samo prikaz trenutnog stanja tehnologije biometrije i prepoznavanja emocija iz glasa, već sadrži i praktični dio implementacije prepoznavanja emocija iz snimke glasa koristeći python programski jezik i njegove biblioteke. Kako se tehnologija nastavlja razvijati, tako će se i naše sposobnosti za interakciju, zaštitu i razumijevanje jedni drugih povećavati, otvarajući nove puteve prema budućnosti u kojoj tehnologija i čovječanstvo idu ruku pod ruku. U konačnici, ovaj rad potvrđuje da, iako smo mi ti koji upravljamo tehnologijom, tehnologija isto tako oblikuje nas. U skladu s tim, i predviđanjima tržišta biometrije glasa, naša budućnost u domeni biometrije i prepoznavanja emocija iz glasa izgleda obećavajuće, s mogućim potencijalom za inovacije koje će nam pružati veću sigurnost.

Popis literature

[1] Anil K. Jain; Patrick Flynn; Arun A. Ross, „Handbook of Biometrics“, Springer, 2007.

[2] Julian Ashbourn, „Biometrics Advanced identity verification“, Springer, 2000.

[3] Anil K. Jain; Karthik Nandakumar; Arun A. Ross, „Introduction to Biometrics“, Springer, 2011.

[4] Anil K. Jain; Arun Ross; Salil Prabhakar, „An Introduction to Biometric Recognition“, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 14, NO. 1, 2004.

[5] „What is a voice biometrics“ (03.2023), Mobbel [Na internetu] Dostupno: Mobbeel, https://www.mobbeel.com/wp-content/uploads/2023/03/What-is-voice-biometrics.pdf

[6] „Voice biometrics“ (2000.), Judith A. Markowitz [Na internetu] Dostupno: ACM, https://dl.acm.org/doi/pdf/10.1145/348941.348995

[7] „Biometric Voice Recognition – Everything You Should Know“ (bez dat.) Imageware [Na internetu] Dostupno: Imageware, https://imageware.io/biometric-voice-recognition/

[8] „Speech Emotion Recognition Project using Machine Learning“ (08.02.2024.) Param Raval [Na internetu] Dostupno: Projectpro, https://www.projectpro.io/article/speech-emotion-recognition-project-using-machine-learning/573

[9] „Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots“ (2017.) Eduard Frant i ostali [Na internetu] Dostupno: Romjist, https://www.romjist.ro/full-texts/paper562.pdf

[10] Lee, WS., Roh, YW., Kim, DJ., Kim, JH., Hong, KS. (2008). Speech Emotion Recognition Using Spectral Entropy. In: Xiong, C., Liu, H., Huang, Y., Xiong, Y. (eds) Intelligent Robotics and Applications. ICIRA, 2008.

[11] „What is the KNN algorithm?“ (bez dat.), IBM [Na internetu] Dostupno: IBM, https://www.ibm.com/topics/knn

[12] „What is a convolutional neural network (CNN)?“ (bez dat.), Lev Craig [Na internetu] Dostupno: Techtarget,https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network

[13] „What are recurrent neural networks?“ (bez dat.), IBM [Na internetu] Dostupno: IBM, https://www.ibm.com/topics/recurrent-neural-networks

[14] „What is deep learning“ (bez dat.), IBM [Na internetu] Dostupno: IBM, https://www.ibm.com/topics/deep-learning

[15] „What are support vector machines (SVMs)?“ (bez dat.), IBM [Na internetu] Dostupno: IBM, https://www.ibm.com/topics/support-vector-machine

[16] „An ongoing review of speech emotion recognition“ (2023), Javier de Lope, Manuel Graña [Na internetu] Dostupno: Sciencedirect, https://www.sciencedirect.com/science/article/pii/S0925231223000103

[17] „Emotion recognition from speech: a review“ (2012), Shashidhar G., Koolagudi K., Sreenivasa Rao [Na internetu] Dostupno: Brooklyn, https://www.sci.brooklyn.cuny.edu/~levitan/nlp-psych/papers/koolagudi12.pdf

[18] Ajith Abraham; Anu Bajaj; Niketa Gandhi, „Innovations in Bio-Inspired Computing and Applications“, Springer, 2022.

[19] „Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer“ (07.07.2023), Rizwan Ullah i ostali [Na internetu] Dostupno: MDPI, https://www.mdpi.com/1424-8220/23/13/6212

[20] „Speech Emotion Recognition Using Deep Learning Techniques: A Review“ (19.08.2019), Ruhul Amin i ostali, [Na internetu] Dostupno: Ieeexplore, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8805181

[21] „Speech Emotion Recognition Using Audio Matching“ (25.11.2022), Iti Chaturvedi i ostali [Na internetu] Dostupno: MDPI, https://www.mdpi.com/2079-9292/11/23/3943

[22] Plutchik Robert, The nature of emotions, American Scientist 89, 2001.

[23] „Speech emotion recognition using machine learning — A systematic review“ (20.11.2023) Samaneh Madanian i ostali, [Na internetu] Dostupno: Sciencedirect, https://www.sciencedirect.com/science/article/pii/S2667305323000911

[24] „Classifying emotions using audio recordings and Python“ (25.02.2021), Tal Baram [Na internetu] Distupno: Towards dana science, https://towardsdatascience.com/classifying-emotions-using-audio-recordings-and-python-434e748a95eb

Popis slika

Popis tablica

Prilozi

[1] Source code aplikacije: https://github.com/DigitalDevelooper/SER
[2] RAVDESS baza podataka - https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio