Raspoznavanje emocija iz zvučnih snimki govora

Zagoršćak, Martin

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij

RASPOZNAVANJE EMOCIJA IZ ZVUČNIH SNIMKI GOVORA

Diplomski rad

Martin Zagoršćak

Osijek, 2022.

Sadržaj

1. UVOD

Raspoznavanje je osnovna vještina identificiranja na temelju stečenog znanja i iskustva. Živa bića kontinuirano i nesvjesno raspoznaju stvari i okolinu u kojoj se nalaze te konstantno uče raspoznavati nove stvari i okoline, time unaprjeđujući vlastitu vještinu raspoznavanja. Dolaskom uređaja i tehnologija koja povećavaju kvalitetu života ljudskim bićima, pojavila se potreba za unaprjeđenjem uporabljivosti tehnologija te se razvila znanstvena disciplina koja se bavi komunikacijom između čovjeka i računala (engl. *Human-Computer Interaction*). Cilj te znanstvene discipline je prilagoditi sustav korisniku. Uz pojednostavljene upute za korištenje sustava (kako bi se uporabljivost sustava iskoristila maksimalno), vizualnog dizajna sustava i odziva sustava na korisničku interakciju, bitan faktor je raspoznavanje tipa, odnosno profila, korisnika koji ga koristi kako bi se sustav znao prilagoditi korisniku. Za što točniju prilagodbu sustava korisniku, sustav bi morao prepoznati njegovu kulturu, kognitivne sposobnosti, korisničko iskustvo i korisnikove emocije. Emocije je načelno moguće raspoznati iz govora tijela, izraza lica, konteksta, ljudskog govora itd. Problem nastaje kada računalno treba raspoznati emociju kod čovjeka. Rješenje navedenog problema ostvaruje se pomoću strojnog učenja (engl. *machine learning*), za koje danas postoje razvijeni postupci i alati koji ipak i dalje nisu blizu savršenog. Računalno raspoznavanje emocija eventualno će se približiti tomu kada bude u mogućnosti raspoznavati emocije na razini čovjeka koji može istovremeno iz više izvora ljudske komunikacije raspoznati emociju, dob, kulturu i slično.

Cilj ovog diplomskog rada je prikazati problem računalnog raspoznavanja emocija i njegove glavne značajke. Fokus je postavljen na raspoznavanje emocija iz zvučnih zapisa govora. Rad je raspoređen u 4 poglavlja od kojih prvo predstavlja uvod u dani problem. U drugom poglavlju opisan je model emocija te postupci i primjena raspoznavanja emocija. Također, dana su postojeća rješenja na Android platformi. Treće poglavlje donosi usporedbu različitih postupaka strojnog učenja, definirane zahtjeve programskog rješenja, prikazane slučajeve korištenja te mogućnost ugradnje modela strojnog učenja u Android aplikaciju. Posljednje poglavlje predstavlja zaključak, odnosno osvrt, na problem računalnog raspoznavanja emocije te implementaciju rješenja za navedeni problem.

2. RASPOZNAVANJE EMOCIJA

U međuljudskoj komunikaciji bitnu ulogu igraju emocije i njihova interpretacija. Prema [1], u jednostavnom dijalogu između dvije osobe komunikacija se izvodi verbalno, govorom tijela, izrazom lica, tonom glasa i ostalim neverbalnim izričajima. Na sve navedene izvore informacija koje dolaze do sugovornika utječe trenutno emocionalno stanje, tako da za navedeni primjer jednostavnog dijaloga, na temelju trenutne emocije govornik bira riječi, postavlja svoje tijelo u određeni položaj, stvara izraz lica te prilagođava ton glasa. Čovjek kao sugovornik često ne nailazi na prepreke pri interpretaciji emocija i informacija. Ukoliko dođe do pogrešne interpretacije emocija, a uz to i informacija, navedena se situacija naziva nesporazum. Za razliku od ljudskog, računalno raspoznavanje emocija susreće se još brojnijim problemima. Na te probleme utječu mnogi faktori, primjerice za zvuk, razni šumovi, za fotografiju prekriveno lice rukom ili za tekstualni oblik, kontekst. Cilj sustava koji sadrže mehanizam raspoznavanja emocija je pravodobno i precizno prepoznati emociju te reagirati na nju. Time kvaliteta programske podrške raste i korisnici sustava su zadovoljni.

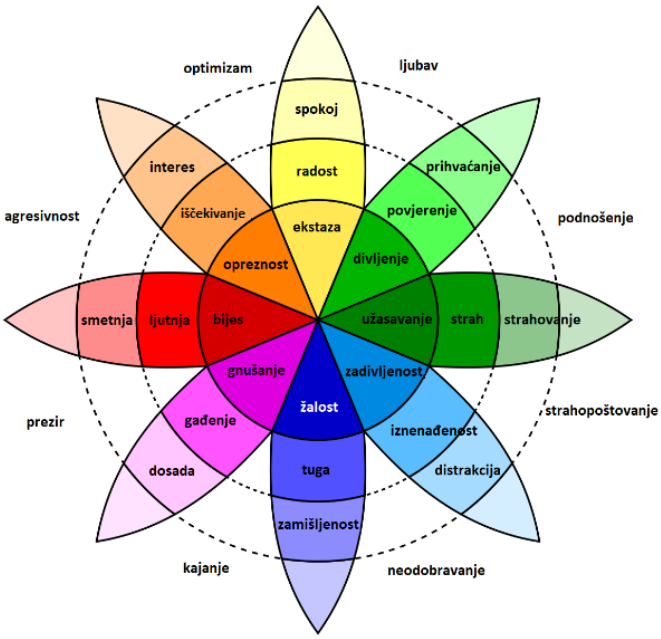
2.1. Model emocija

Sve teorije o emocijama razvile su se unutar povijesnog konteksta. Značajne osobe koje su doprinijele današnjem pojmu emocije su Charles Darwin, William James i Sigmund Freud. Pojam emocije se opisuje kao spontana reakcija na vanjski ili unutrašnji podražaj živog bića. Konkretnija definicija, prema [2] glasi da emocije potječu iz subkortikalne regije i ventromedijalnog prefrontalnog korteksa mozga, koji uzrokuju biokemijske reakcije i mijenjaju ljudsko fizičko stanje. Vanjskim podražajem smatra se bilo kakva interakcija s okolinom, dok se unutrašnjim podražajem smatra bilo kakav događaj unutar tijela, npr. razmišljanje ili glad. Često dolazi do nerazlikovanja između emocija i osjećaja. Prema [3], osjećaji se od emocija razlikuju u trajanju. Prvo dolazi do emocija te onda nakon što ljudsko tijelo osjeti promjene u fizičkom stanju dolazi do osjećaja. Uz emocije i osjećaje veže se i pojam raspoloženja, koje predstavlja skup osjećaja u nekom vremenskom razdoblju.

Psiholog Robert Plutchik razvio je Kotač Emocija (engl. *Emotion Wheel*) (Sl. 2.1), na kojemu prikazuje 8 osnovnih emocija: radost, povjerenje, strah, iznenađenje, tuga, iščekivanje, ljutnja i gađenje. Emocija na kotaču su poslagane kružno zbog sličnosti susjednih emocija i tako da svaka osnovna emocija ima svoju polarnu suprotnost: radost i tuga, strah i ljutnja, iščekivanje i iznenađenje te gađenje i povjerenje. Osim redoslijeda emocija, prikazani su i intenziteti emocija,

tako za primjer radosti, slabiji intenzitet predstavlja spokoj dok za jači intenzitet predstavlja ushićenost. Plutchik je razvio 10 postulata na kojemu se temelji njegova teorija o emocijama, koji prema [4] glase:

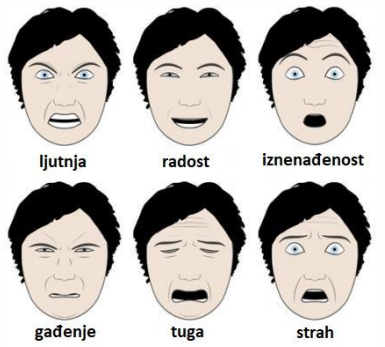
1. Ljudska bića i životinje doživljavaju osnovne emocije na sličan način.
2. Emocije su rezultat evolucije. Bile su prisutne kod životinja i prije nego što su evoluirali majmuni.
3. Emocije su evoluirale u svrhu povećanja šansi preživljavanja u okolini.
4. Unatoč što postoji nekoliko vrsta emocija koje su prisutne u različitim vrstama, postoje prototipski obrasci emocija koji se mogu identificirati.
5. Postoji relativno mali broj prototipa, osnovnih emocija koje se mogu identificirati.
6. Sve ostale emocije se smatraju kombinacijom osnovnih emocija.
7. Osnovne emocije su idealizirana stanja koja se opisuju u smislu njihovih karakteristika.
8. Primarne emocije su kategorizirane kao parovi polarnih suprotnosti.
9. Sve emocije imaju različite stupnjeve sličnosti s drugim emocijama.
10. Svaka emocija može varirati u svojoj razini intenziteta.



Slika 2.1 Plutchikov kotač emocija, izrađeno prema [4]

2.2. Postupci raspoznavanja emocija

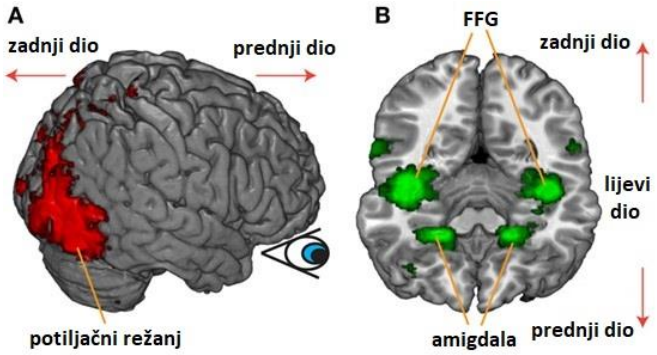
Emocije u ljudskoj interakciji mogu se prepoznati vizualno, auditivno i tekstualno. Postupak vizualnog raspoznavanja emocija temelji se na praćenju izraza lica (Sl. 2.2). U razgovoru se ljudi fokusiraju na oči ili na usta, zbog toga što daju preciznije informacije o kojoj se trenutno emociji radi. Osim očiju ili usta, tjelesno izražavanje emocije se može primijetiti položajem i kretnji ruku, nogu, obrva, obraza, nosa...



Slika 2.2 Prikaz 6 osnovnih emocija izrazom lica, izrađeno prema [5]

Prema [6], za proces vizualnog raspoznavanja emocija su zaslužni (Sl. 2.3):

- vidni živci, očne jabučice
- *fusiform gyrus* (FFG) - unutar potiljačnog režnja (engl. *occipital lobe*), obrađuje informacije ljudskog vida
- amigdala – manja regija koja se nalazi na prednjem dijelu temporalnog režnja, obrađuje emocije



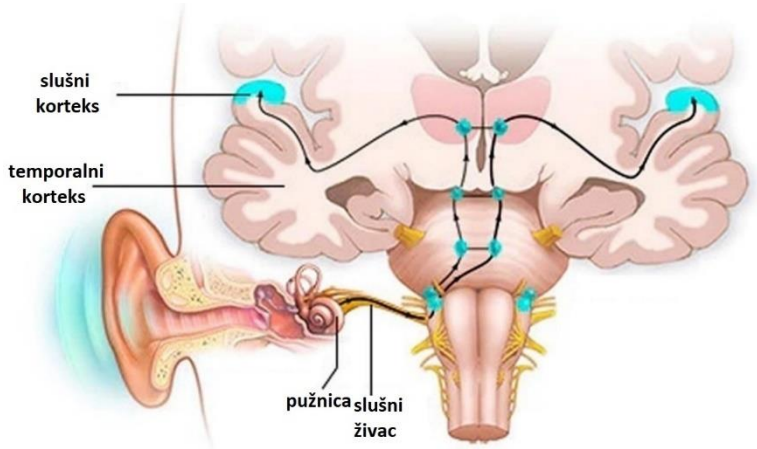
Slika 2.3 Prikaz lokacije potiljačnog režnja, FFG-a i amigdale na mozgu, izrađeno prema [6]

Postupak auditivnog raspoznavanja emocija temelji se na govoru govornika koji prikazuje emociju zvučno što se opisuje visinom (frekvencija zvučnog vala), jačinom (amplituda zvučnog vala) te

trajanjem (vrijeme izvođenja) glasa i lingvističkim značajkama. Za primjer ushićenosti, osoba je glasnija, brža s izgovorom, izgovara sa snažnom visokofrekventnom energijom i širim rasponom tona. Dok za osobu kod koje prevladava tuga, njen izgovor je sporiji, nižeg tona i s manje energije.

Prema [7], za proces auditivnog raspoznavanja emocija su zaslužni (Sl. 2.4):

- uši i slušni živci
- slušni korteks (engl. *auditory cortex*) – nalazi se u temporalnom režnju, obrađuje informacije zvuka
- također amigdala koja komunicira sa slušnim korteksom

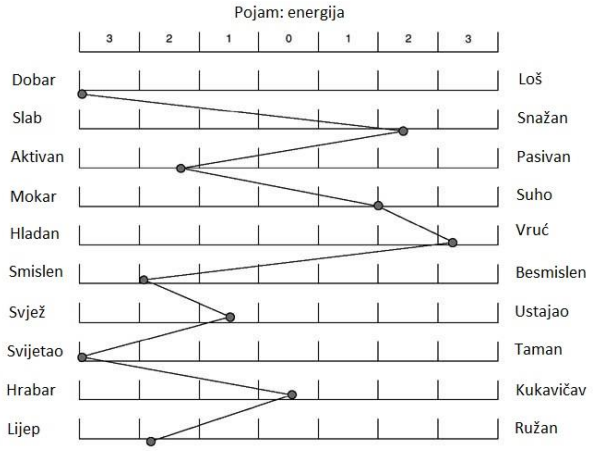


Slika 2.4 Prikaz slušnog sustava povezanim sa slušnim korteksom, izrađeno prema [7]

Iako pri čitanju tekstualne poruke nedostaju izrazi lica, akustične značajke govora, vizualni govor tijela i slično, ljudi i dalje mogu raspoznati emociju ili raspoloženje osobe. Psiholog Charles E. Osgood, prema [8, str. 9], iznio je teoriju Semantičkog Razlikovanja (engl. *Semantic Differentiation*) koja se bavi pridruživanjem emocionalnog značenja riječima na 3 faktora:

- procjena (engl. *evaluation*) na razini dobar-loš
- moć (engl. *potency*) na razini snažan-slab
- aktivnost (engl. *activity*) na razini aktivan-pasivan.

Može se zaključiti kako čovjek na temelju iskustva i uporabe riječi u određenim kontekstima subjektivno dodjeljuje emociju značenju neke riječi. Na slici 2.5 prikazan je primjer semantičkog ocjenjivanja riječi energija.



Slika 2.5 Prikaz semantičkog ocjenjivanja riječi „energija“ uz dodatne faktore, izrađeno prema [9, str. 812]

Socijalni psiholog Michael Kraus, prema [10], eksperimentom u kojem je sudjelovalo više od 1800 ljudi iz Sjedinjenih Američkih Država, zaključio je da se emocije najučinkovitije raspoznaju slušanjem. Lošiji rezultati su dobiveni vizualnim raspoznavanjem i kombinacijom vizualnog i slušnog raspoznavanja emocija. Osim 3 navedena slučaja raspoznavanja emocija, sudionici su slušali računalni glas koji čita transkript ljudskog dijaloga i pokušali raspoznati emocije što se pokazalo najgorim slučajem raspoznavanja. Kraus smatra da više informacija ne znači uvijek veću preciznost i u kognitivnom smislu smatra da je kompleksno pratiti 2 izvora informacija istovremeno. Također, navodi razlog da ljudi često prikrivaju pravu emociju odglumljenim izrazom lica.

2.3. Računalno raspoznavanje emocija i njegova primjena

Tehnologija računalnog raspoznavanja emocija danas je naveliko korištena među brojnim tvrtkama i ustanovama. Kao što je već rečeno, cilj računalnih sustava koji ovise o korisničkoj interakciji je prilagoditi se prema potrebama korisnika. Sustav koji je sposoban raspoznati emociju u danoj okolini podiže razinu uporabljivosti i učinkovitosti od koje profitiraju i pružatelj proizvoda ili usluge i korisnik. Iako se danas računalno raspoznavanje emocija naveliko koristi, nije ni približno savršeno, nailazi na brojne prepreke i izazove. Velik izazov stvara raznolikost kultura i rasa. Pri vizualnom raspoznavanju emocija, prema [11], problem mogu stvarati biološke osobine kao što su oblik i gustoća obrva, oblik usta, oblik obraza i slično, za prepoznavanje iste emocije između različitih kultura. Pri auditivnom raspoznavanju emocija, problem predstavljaju različite dubine glasa, primjerice za muški rod, istočni Azijati višu dubinu glasa dok osobe crne rase imaju nižu dubinu glasa. Osim tjelesnih osobina, prepreku predstavljaju različite osobnosti korisnika.

Tako pri računalnom raspoznavanju emocije iz teksta, prema [12], sarkazam i kontekst mogu stvarati prepreku pri točnosti klasifikacije emocije. Osim već navedenih izazova, problem stvaraju i različitost u dobi, brzini govora i energičnosti govornika. Idealni cilj bi bio dostići stopostotnu točnost klasifikacije emocije, ali je to gotovo nemoguće zbog navedenih prepreka i činjenice da se izrazi lica i dubina glasa mogu odglumiti.

Područja u kojima raspoznavanje emocija igra ulogu, prema [13, 14]:

- Zdravstvo – prioritiziranje pacijenata u čekaonici, promatranje stanja pacijenata tijekom liječenja, otkrivanje mentalnih bolesti kao što je depresija, prevencija samoubojstva te procjena kada je pacijentu potreban lijek.
- Automobilska industrija - povećanje razine fokusiranosti vozača na vožnju ili detekcija umora kod vozača.
- Pomoć u području ljudskih resursa – procjena kandidatove zainteresiranosti i motiviranosti tijekom razgovora, raspoznavanje nezainteresiranih kandidata tijekom intervjua, praćenje raspoloženja zaposlenika u svrhu poboljšanja kvalitete radnog okruženja.
- Kod osoba različitih sposobnosti - pomoć autističnim osobama, osobama s poremećajima i starijim osobama pri interpretaciji emocija drugih sudionika u okolini u kojoj se nalaze.
- Javna sigurnost – kontrole na graničnim prijelazima, poboljšanje pri detektiranju laži, procjena sigurnosti javnih okruženja u svrhu prevencije potencijalnog terorizma.
- Personalizirane usluge – prilagodba pametnog okruženja, prilagođena ponuda glazbe, kulturnih događaja, oglasa na temelju profila korisnika.
- Edukacija – praćenje emocionalnih stanja studenata pri kojima su zainteresiranost ili angažiranost i zbunjenosti najzastupljeniji u cilju poboljšanja okoline učenja. Moguće je i procijeniti kvalitetu edukacijskih resursa i na temelju procjene odabrati resurse za efikasnije i zanimljivije učenje.
- Zabava – puštanje zabavnog sadržaja poput glazbe ovisno o korisnikovom emocionalnom stanju.
- Razvoj programske podrške – prepoznavanje korelacija između emocija razvojnih programera i produktivnosti ili kvalitete koda. Osim navedenog, korisno je i pri praćenju razlike između rada od kuće ili u uredu.
- Poboljšanje programske podrške – praćenje emocija prilikom testiranja uporabljivosti aplikacije gdje se provode testovi poput testa prvog dojma, test korištenja implementiranih

funkcionalnosti, testa slobodnih interakcija (engl. *free interaction*) i testa usporedbe dviju verzija aplikacije.

- Video igre – detaljna analiza stvarnog vremena emocija tijekom korištenja. Raspoznavanjem emocija igrača moguće je povećati razinu zainteresiranost i produljiti vrijeme igre prilagodbom podražaja koji utječu na igrača.

Prema [15], dobri rezultati su postignuti u edukacijskim svrhama. U Hong Kongu, za vrijeme pandemije platforma *4 Little Trees* prati lica učenika te pomaže učiteljima „pročitati" stanje razreda za vrijeme predavanja i učenja. Dok učenici pišu zadaću ili ispite, sustav mjeri mišićne točke na licu pomoću kamera na računalima ili tabletima i na temelju njih prepoznaje emocije sreće, tuge, ljutnje, iznenađenja i straha. Osim raspoznavanja emocija, sustav prati koliko dugo je potrebno učenicima da odgovore na pitanja te prati njihove ocijene i napredak. Također, generira izvješća o učenicima o njihovim vrlinama, manama i razinama motivacije. Sustav se prilagođava svakom učeniku ponaosob. Cilj je i povećati motivaciju učenika za radom i napretkom, tako sustav nudi učenje na razini igre gdje učenici skupljaju novčiće. Učenici postižu 10% bolje rezultate na ispitima koristeći ovu platformu. Što se tiče privatnosti učenika, autori smatraju da je transparentnost ključ za održavanjem privatnosti. Sustav treba tražiti suglasnost roditelja ili skrbnika za prikupljanje podataka i u koje će se sve svrhe koristiti ti podaci. Kao problem navode za raspoznavanje tamnijih rasa zbog učenja sustava prema većinom bijeloj rasi, ali u Hong Kongu sustav postiže točnost od 85%. Za raspoznavanje jednostavnih emocija sreće i tuge postiže 90%-tu točnost, dok za složenije emocije kao što su entuzijazam ili anksioznost, sustav teže raspoznaje.

Iako se računalno raspoznavanje emocija koristi u brojnim područjima, zbog ranije navedenih izazova koji predstavljaju problem za preciznost sustava pri raspoznavanju emocija, sustavi mogu pokazati diskriminaciju prema određenoj populaciji. Na primjer, za crnu rasu, sustavi češće pretpostavljaju emociju ljutnje nego kod ostalih rasa. Za bolju točnost sustava potrebno bi bilo uzeti u obzir različite rase i kulture kako ne bi došlo do prenaučenosti sustava na jednu rasu i u konačnosti na diskriminaciju. Koliko god sustav raspoznavanja emocija pružao korisniku pametnu uslugu, prema [16], moguće je kod korisnika probuditi osjećaj nesigurnosti i ugroženosti te mu promijeniti ponašanje na temelju njegove izloženosti sustavu i narušavanju njegove privatnosti. Navedene nuspojave događaju se kada korisnik sazna što sve sustav o njemu zna, tj. koliko je sustav upoznat i prilagođen korisničkom profilu. Računalni sustav koji prepoznaje emocije ne bi trebao donositi odluke koje bi mijenjale ili znatno utjecale na korisnikov život te bi sustav morao zatražiti pristanak korisnika za rezultat koji sustav pruža na temelju njegove klasificirane emocije.

2.4. Postupci računalnog raspoznavanja emocija

Raspoznavanje emocija na licu (engl. *Face Emotion Recognition (FER)*) je postupak klasifikacije emocija na osnovi videozapisa ili fotografije lica. Prema [17], fotografiji lica je prvo potrebno ukloniti pozadinu, izdvojiti vektor značajki koji sadrži informacije o ključnim točkama lica i provući kroz primjerice, konvolucijsku neuronsku mrežu (engl. *Convolutional Neural Network (CNN)*) koja u konačnici daje rezultat klase emocije. Važne točke vektora lica su nos, usne, čelo, oči, uši, brada.

Govorno raspoznavanje emocija (engl. *Speech Emotion Recognition (SER)*) je postupak nadziranog učenja kojemu je cilj raspoznati emociju na temelju govora. Zadatak SER-a je vrlo zahtjevan zbog ranije navedenih izazova (različitost u dobi, kulturi, rasi i tako dalje). Iz ljudskog govora moguće je na 2 načina raspoznati o kojoj emociji se radi, putem leksičkih značajki i putem akustičkih značajki. Obje inačice problema predstavljaju višeklasne probleme klasifikacije te se za njihovo rješavanja često koriste algoritmi strojnog učenja. U sljedećim potpoglavljima je detaljno objašnjen pojam raspoznavanja emocija kao klasifikacijski problem s pripadajućim mjerama procjene kvalitete te postupak računalnog raspoznavanja emocije, od prikupljanja i predobrade podataka, izdvajanja značajki i često korištenih algoritama strojnog učenja za rješavanje navedenog problema.

2.4.1. Raspoznavanje emocija kao klasifikacijski problem

Klasifikacija je oblik nadziranog strojnog učenja u kojemu su primjercima skupa podataka dodijeljene unaprijed poznate oznake klasa, a na temelju poznatih oznaka novim primjercima potrebno je dodijeliti jednu među njima. Klasifikacija se javlja u dva oblika, binarna klasifikacija ili višeklasna (engl. *multiclass*) klasifikacija. Algoritmi za provedbu procesa klasifikacije nazivaju se klasifikatorima, a oni se dijele na dvije grupe, lijeni (engl. *lazy*) i marljivi (engl. *eager*). Lijeni klasifikatori pohrane podatke za treniranje i odgađaju obradu podataka sve dok se ne pojavi testna instanca kojoj se treba dodijeliti oznaka klase. Za razliku od lijenih, marljivi klasifikatori kreiraju klasifikacijski model na temelju podataka za treniranje i kada je potrebno odrediti klasu nove instance, ona se provodi strukturiranim modelom. Značajna razlika se ističe u vremenu treniranja (manja kod lijenih klasifikatora) i u vremenu predviđanja (manja kod marljivih klasifikatora). Treniranju klasifikatora prethode prikupljanje i predobrada podataka. Predobradom podataka smatra se pretvaranje podataka iz sirovog oblika u formatirani oblik kojeg analitičar podataka (engl. *data analyst*) može obrađivati i rukovati njime.

Za analizu performansi istreniranog klasifikatora koriste se mjere kvalitete poput:

- matrice zabune (engl. *confusion matrix*) – kvadratna matrica u kojoj retci predstavljaju stvarnu klasu dok stupci predstavljaju klase koje je klasifikator raspoznao. Pomoću matrice zabune moguće je uočiti s kojim klasama klasifikator ima poteškoća pri raspoznavanju i koje klase s lakoćom raspoznaje. Na temelju matrice zabune moguće je izračunati druge mjere kvalitete koje će u nastavku biti izvedene. Elementi matrice su:
  - stvarno pozitivni (engl. *True Positives* (TP)),
  - lažno negativni (engl. *False Negatives* (FN)),
  - lažno pozitivni (engl. *False Positives* (FP)),
  - stvarno negativni (engl. *True Negatives* (TN)).
- točnosti (engl. *accuracy*) – mjera točno raspoznatih instanci izražena u postotcima, računa se kao

$$A = \frac{TP + TN}{TP + FP + TN + FN}. \quad (2\text{-}1)$$

- preciznosti (engl. *precision*) – mjera točnosti raspoznavanja klase da je ona zapravo stvarna, također izraženo u postotcima i računa se kao

$$P = \frac{TP}{TP + FP}. \quad (2\text{-}2)$$

- odziva (engl. *recall*) – mjera koja predstavlja koliko je stvarno pozitivnih raspoznato od ukupnog broja stvarnih instanci te klase, također izraženo u postotcima te se računa se kao

$$R = \frac{TP}{TP + FN}. \quad (2\text{-}3)$$

- mjere F1 (engl. *F1-score*) – mjera za određivanje točnosti testa gdje rezultat može biti u intervalu [0, 1], računa se kao harmonijska sredina preciznosti i odziva

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}. \quad (2\text{-}4)$$

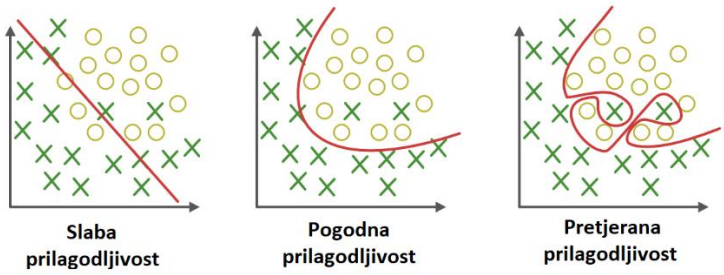
- logaritamski gubitak (engl. *log loss*) – mjera gubitka u klasifikaciji koja se računa kao

$$L_{log}(y, p) = -(y ln(p) + (1 - y) \ln(1 - p)). \quad (2\text{-}5)$$

  gdje y predstavlja pravu vrijednost instance koja može biti 0 ili 1, a p vjerojatnost instance da pripada točnoj klasi prema klasifikatoru.

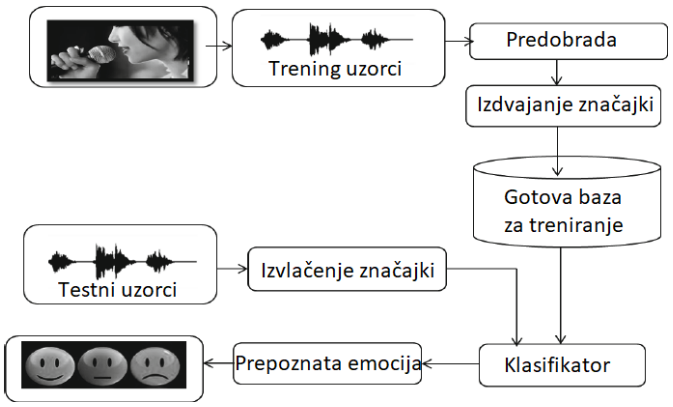
Usporedbom performansi tijekom treniranja i testiranja modela strojnog učenja moguće je prepoznati dva potencijalna problema za loše rezultate testiranja, pretjeranu prilagodljivost (engl. *overfitting*) i slabu prilagodljivost (engl. *underfitting*) (Sl. 2.6). Prema [18], pretjerana

prilagodljivost predstavlja scenarij kada je model previše privržen podacima za trening, tj. da je snažno usmjeren na detalje i šum unutar podataka. Najčešći uzrok pretjerane prilagodljivosti je prevelika kompleksnost modela. Postupci za smanjenje pretjerane prilagodljivosti su dodavanje dodatnih podataka za trening, pojednostavljivanje modela, uklanjanje šuma u podacima. Suprotnost pretjerane prilagodljivosti je slaba prilagodljivost koja također može rezultirati lošim predviđanjima pri testiranju. Prema [18], slaba prilagodljivost je scenarij u kojemu je model prejednostavan, tj. ne može naučiti osnovnu strukturu podataka. Najčešći uzrok slabe prilagodljivosti je manjak trening podataka i izgradnja linearnog modela nad nelinearnim podacima. Slaba prilagodljivost se može riješiti postupcima povećanja kompleksnosti modela, inženjerstvom značajki (engl. *feature engineering*) ili smanjenjem ograničenosti modela.



Slika 2.6 Prikaz razina prilagodljivosti, izrađeno prema [18]

Proces raspoznavanja emocija iz zvučnih snimki govora prikazan na slici 2.7 započinje prikupljanjem podataka za treniranje modela strojnog učenja. Baze podataka za govorno raspoznavanje emocija se kreiraju snimanjem glumljenih ili snimanjem stvarnih emocija i pohranjivanjem (najčešće) u .wav formatu. Kako se ne bi ručno snimali audiozapisi, putem Interneta su dostupne gotove baze podataka. Za svaku instancu iz baze odvija se potrebna predobrada, potom se izdvajaju željene značajke s kojima će se klasifikacija provoditi. Nakon što se iz baze podataka audiozapisa kreirala baza podataka s izdvojenim značajkama i oznakom klase, nova baza je spremna za dijeljenje podataka na skupove za treniranje i testiranje. Podaci za trening utječu na parametre modela strojnog učenja, dok podaci za testiranje služe za izradu informacija o kvaliteti modela. Kako bi se ubuduće koristio naučeni model, on se pohranjuje i pretvara u konačni modul koji se integrira u sustav ili objavljuje na online servis. Za nove audiozapise kojima se tek treba odrediti emocija, izdvajaju se značajke i prosljeđuju se naučenom modelu koji u konačnici vraća oznaku emocije. Detaljniji opisi koraka eksperimenta se nalaze u sljedećim potpoglavljima.



Slika 2.7 Struktura projekta strojnog učenja za govorno raspoznavanje emocija, izrađeno prema [19, str. 337]

2.4.2. Prikupljanje podataka i dostupni podatkovni skupovi

Baze podataka (podatkovni skupovi) su iznimno bitne u izradi sustava za raspoznavanje emocija na temelju glasovnog zapisa zbog toga što se izgradnja modela kod klasifikacije oslanja na označene podatke. Baze podataka za ovaj problem, prema [20] moguće je podijeliti u 3 skupine:

- Baza s glumljenim emocijama – profesionalni ili amaterski glumci čitaju pripremljene rečenice unutar zvučno izoliranih studija u kojima iznose emocije prilikom čitanja rečenica. Manu ovakve baze predstavlja manjak stvarnosti unutar emocija koje bi se pojavljivale u pravim životnim situacijama te time ponekad glumljene emocije mogu biti pretjerane.
- Baza s emocijama prirodnog govora – izvore predstavljaju prave životne situacije kao na primjer: televizijske emisije, snimci iz pozivnih centara, radio razgovori i slično. Takve audiozapise je teško za prikupiti zbog etičkih i pravnih prepreka.
- Baza s emocijama izazvanog govora – glumci se postavljaju u simuliranu emocionalnu situaciju gdje izvode monologe ili dijaloge i pri tome su emocije sličnije stvarnim emocijama za razliku od glumljenih emocija koje se temelje na čitanju rečenica.

Odabir podataka za treniranje modela strojnog učenja je vrlo bitan zbog toga što se govorno raspoznavanje emocija temelji na nadziranom učenju, odnosno na označenim (engl. *labeled*) podacima. Ukoliko su podaci nekvalitetni ili nepotpuni, može doći do pogreške pri klasificiranju emocije. Za razvoj aplikacije kojom će stvarni korisnik upravljati, idealno bi bilo da model strojnog učenja bude učen na temelju prirodnih audiozapisa. Nažalost, većina besplatno dostupnih baza podataka su glumljene. Glumljene baze podataka koje se temelje na izgovoru iste rečenice kroz više emocija su pogodne za znanstvena istraživanja akustičnih značajki, zbog toga što se

model strojnog učenja neće temeljiti na leksičkim značajkama, već na akustičnim. Među brojnim dostupnim skupovima podataka, ističu se:

- **Acted Emotional Speech Dynamic Database (AESDD)** – javno dostupna baza podataka u svrhu govornog raspoznavanja emocija. Iskazi su na grčkom jeziku. Audiozapise su snimila 2 glumca i 3 glumice. Audiozapisi su označeni (engl. *labeled*) emocijama sreće, tuge, gađenja, straha i ljutnje. Svaki audiozapis je imenovan formom: „xAA (B)“, gdje x označava prvo slovo emocije, AA označava identifikaciju izgovorene rečenice, B označava identifikaciju govornika [21].
- **Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)** – baza podataka od 7442 audiozapisa 48 glumaca i 43 glumice u dobi od 20 do 74 godine različitih kultura. Audiozapisi su predstavljeni emocijama sreće, tuge, gađenja, straha, ljutnje i neutralnog osjećaja. Audiozapisi su imenovani formom, primjer: „1037_DFA_ANG_XX“, gdje 1037 predstavlja identifikaciju glumca, DFA identifikaciju izgovorene rečenice, ANG emociju ljutnje, XX intenzitet [22].
- **EMOVO** – prva audio baza glumljenih emocija na talijanskom jeziku. Glasovni zapisi su prikupljeni od 3 glumice i 3 glumca. Glasovni zapisi uključuju emocije gađenja, sreće, straha, ljutnje, iznenađenja, tuge i neutralnog osjećaja. Svaki glumac je iskazao svaku navedenu emociju na 14 različitih fraza, što rezultira s ukupno 588 glasovnih zapisa. Imena zapisa su imenovani formom, primjer: „neu-m1-b1“, gdje neu predstavlja neutralan osjećaj, m1 identifikator glumca i b1 identifikator rečenice [23].
- **Berlin Database of Emotional Speech (Emo-DB)** – baza podataka audiozapisa iskazanih emocijom koja je na njemačkom jeziku. Sadrži oko 500 izjava od 5 glumaca i 5 glumica u dobi od 21 do 35 godina. Iskazane su emocije sreće, tuge, ljutnje, straha, dosade, gađenja i neutralnog osjećaja. Audiozapisi su imenovanim formom od 7 znakova: „aaBBBcD“ gdje aa predstavlja identifikaciju govornika, BBB identifikaciju rečenice, c prvo slovo emocije na njemačkom jeziku, D verziju, ukoliko ih ima više [24].
- **The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** – baza podataka emocionalnog govora i pjesama koja sadrži 7356 datoteka (koristit će se samo govorni audiozapisi). Audiozapise je izvelo 12 glumaca i 12 glumica. Govori su iskazani srećom, tugom, ljutnjom, strahom, gađenjem, iznenađenjem i neutralnim osjećajem, te su predstavljeni normalnim i jakim intenzitetom. Audiozapisi su imenovani formom: „aa-bb-cc-dd-ee-ff-gg“, gdje aa predstavlja modalitet, bb identifikaciju govora ili

pjesme, cc emociju, dd intenzitet emocije, ee identifikaciju izjave, ff broj ponavljanja, gg identifikaciju glumca [25].

- **Toronto emotional speech set (TESS)** – baza podataka sadrži izgovorenih 200 riječi od 2 glumice u dobi od 26 i 64 godine. Audiozapisi iskazuju emocije sreće, straha, gađenja, ugodnog iznenađenja, ljutnje, tuge i neutralnog osjećaja. Baza sadrži ukupno 2800 audiozapisa. Glumice izgovaraju 200 ciljanih riječi u formi: „Say the word ...“. Forma imenovanja audiozapisa, u 3 dijela: „OAF_youth_happy“, gdje prvi dio OAF predstavlja identifikaciju glumice, drugi dio youth predstavlja izgovorenu riječ, te treći dio emociju [26].

2.4.3. Predobrada podataka

Nakon što je baza podataka pohranjena i audiozapisi označeni emocijom, prije treniranja klasifikatora prethodi predobrada podataka. Prvi korak u pretvorbi sirovog podatka u pripremi podataka za treniranje je uzorkovanje (engl. *sampling*) signala. Rezultat uzorkovanja audio signala je niz ili polje brojeva koji predstavljaju amplitudu signala u određenoj jedinici vremena. Stopom uzorkovanja (engl. *sampling rate*), koja je izražena u Hz, određeno je koliko će uzoraka biti prikupljeno u sekundi. Nadalje je moguće provesti predobradu podataka koja se sastoji od narednih tehnika, prema [20, 27]:

- Prenaglašavanje (engl. *pre-emphasis*) – pojačavanje visokih frekvencija, omogućuje ravnotežu frekvencijskog spektra. Visoke frekvencije uobičajeno su tiše od nižih frekvencija i pomoću viših frekvencija jasnije se razlikuju zvukovi.
- Uokvirenje (engl. *framing*) – dijeljenje kontinuiranog audio signala na segmente fiksne duljine koji se nazivaju okvirima (engl. *frames*). Svrha uokvirivanja leži u tome što se frekvencije u audio signalu mijenjaju tijekom vremena. Primjenom Fourierove transformacije nad audio signalom koji nije uokviren izgubile bi se frekvencijske konture signala tijekom vremena.
- Primjena funkcije prozora (engl. *windowing*) – smanjivanje spektralnog curenja ili bilo kakvog diskontinuiteta (prekida) signala koje nastane prilikom Brze Fourierove Transformacije, a najčešće se koristi Hammingov prozor:

$$H_n = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad (2\text{-}6)$$

za n od 0 do N-1 gdje je N duljina prozora.

- Diskretna Fourierova Transformacija – pretvorba signala iz vremenske u frekvencijsku domenu, njena korisnost proizlazi iz prikaza distribucije frekvencija audio signala koji je pogodan za pripremu ulaznih podataka u klasifikator strojnog učenja. Ova transformacija je temelj izdvajanju spektralnih značajki koje će biti objašnjene u sljedećem potpoglavlju.
- Uklanjanje šuma (engl. *noise reduction*) – eliminacija šuma unutar audio signala kako bi se minimiziralo izobličenje između snimljenog i očekivanog audio zapisa.

2.4.4. Izdvajanje značajki

Glasovne značajke, prema [28], moguće je izdvojiti na dva načina, dijeljenjem glasovnog zapisa na segmente (okvire) i izdvajanjem lokalnih vektora značajki ili izdvojiti globalnu statiku na temelju cijelog glasovnog zapisa. Prednosti globalnih značajka nad lokalnim značajkama se javlja pri kraćem vremenu klasifikacije i treniranja zbog manje dimenzionalnosti podataka. Za globalne značajke se smatra da su efikasnije od lokalnih značajki u slučaju kada se izvodi raspoznavanje između emocija visokih i niskih razina uzbuđenja. Primjerice, za raspoznavanje između ljutnje i tuge globalne značajke se pokazuju učinkovitijima, dok na prepreke nailaze pri raspoznavanju između ljutnje i sreće. Osim navedenog nedostatka globalnih značajki, privremene (engl. *temporal*) informacije se potpuno gube.

Prilikom dizajniranja sustava za raspoznavanje emocija na temelju glasovnog zapisa izrazito je važno odabrati odgovarajuće značajke koje sadrže emocionalan sadržaj. Bitno je naglasiti da odabrane značajke ne smiju ovisiti o govorniku ili o leksičkom sadržaju. Iako su mnoge glasovne značajke istražene, ne postoje najbolje značajke za rješavanje raspoznavanje emocija iz govora te je moguće značajke podijeliti u 4 kategorije (Sl. 2.8), prema [28]:

- neprekidne,
- kvalitativne,
- spektralne
- i TEO (*Teager energy operator*) – temeljene značajke.

Neprekidne značajke mogu predstaviti stanje uzbuđenja govornika iz kojih se može iščitati emocija. Često korištene globalne značajke pri raspoznavanju emocija iz govora su osnovna frekvencija i energija iz kojih se mogu izdvojiti vrijednosti poput medijana, srednje vrijednosti, standardne devijacije, minimuma, maksimuma i slično. Također, brzina govora i formanti mogu se navesti kao često korištene globalne značajke. Kao što je već napomenuto, ovakve globalne značajke nisu učinkovite za raspoznavanje visoko uzbuđenih emocija.

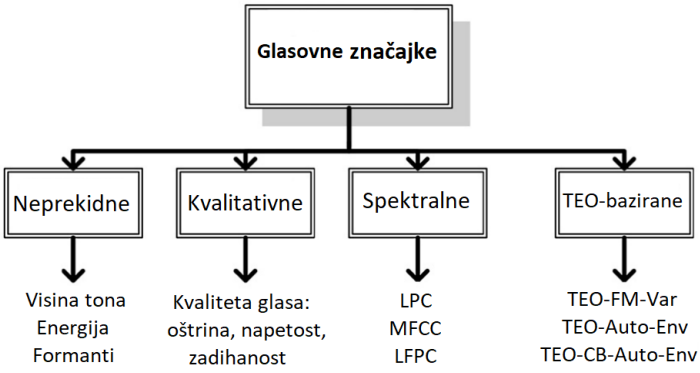
Značajke kvalitete glasa vezane su za emocije snažnog intenziteta koje izrazito utječu na radnje osobe. Pod ovu kategoriju značajki spadaju razina i visina glasa, fraze, fonemi i slično. Zbog opisivanja glasa napetim, oštrim, zadihanim itd., lako dolazi do različitih interpretacija emocija u konačnici. Također, zbog visoke razine kompleksnosti automatiziranog odlučivanja između navedenih opisa kvalitete glasa malo se zna o ulozi kvalitete glasa pri raspoznavanju emocija. Pod ovu kategoriju značajki spadaju razina i visina glasa, fraze, fonemi...

Spektralne značajke se dobivaju tako da se provede Fourierova transformacija koja pretvara signal iz vremenske domene u frekvencijsku domenu. izdvajaju se iz segmenata govora (najčešće duljine 20 do 30 milisekundi) nad kojima je primijenjena metoda prozora te se tako koriste kao lokalne značajke. U ovu kategoriju značajki spadaju MFCC (*Mel Frequency Cepstral Coefficients*), LPCC (*Linear Prediction Cepstral Coefficients*), LFPC (*Log-Frequency Power Coefficients*), GFCC (*Gammatone Frequency Cepstral Coefficients*)...

TEO – temeljene značajke u sustavu raspoznavanja emocija mogu prepoznati govor kao glasan, ljutit, lombardski, jasan ili neutralan. Ova kategorija značajki zasniva se na Teagerovom energijskom operatoru (engl. *Teager energy operator*) koji promatra govor s gledišta energije. Operator su predstavili Teager [29] i Kaiser [30] koji u diskretnoj domeni glasi

$$\psi\{\chi[\eta]\} = \chi^2[\eta] - \chi[\eta - 1]\chi[\eta + 1]. \qquad (2\text{-}7)$$

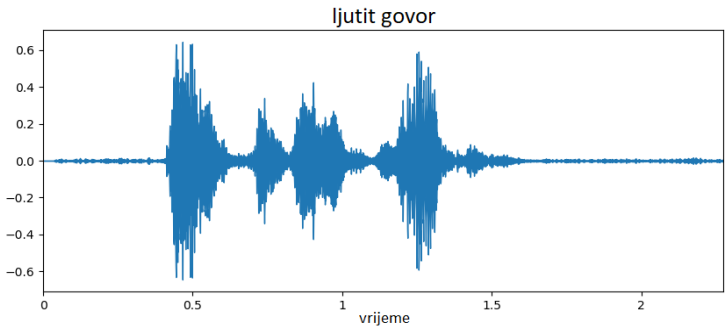
gdje x predstavlja digitalni signal, a n cjelobrojnu varijablu koja predstavlja redni broj uzorka. Prema Teageru, govor je proizveden nelinearno strujanjem zraka iz glasovnog sustava, odnosno kada je govornik pod stresom, njegovi mišići utječu na strujanje zraka iz svojeg glasovnog sustava proizvodeći zvuk.



Slika 2.8 Podjela govornih značajki, izrađeno prema [28]

Prema zaključku iz [28], smatra se da odabir značajki za raspoznavanje emocija iz govora ovisi o potrebnoj zadaći klasifikacije. TEO – temeljene značajke prigodne su za detekciju stresa u govoru. Za razlikovanje između emocija visoke i niske razine uzbuđenja prigodne značajke su osnovna frekvencija i visina glasa te za klasifikaciju jedne od emocija najprigodnije su spektralne značajke koje će se upotrijebiti za praktični dio rada, a opisane su u nastavku.

Audio signal (Sl. 2.9) u vremenskoj domeni pruža informacije o promjeni tlaka zraka tijekom vremena. Nažalost, signal u tom obliku ne pruža dovoljno informacija za prepoznavanje emocija. Poseban značaj ovom projektu pružaju Mel spektrogram i MFCC značajke koje će u nastavku biti opisane.

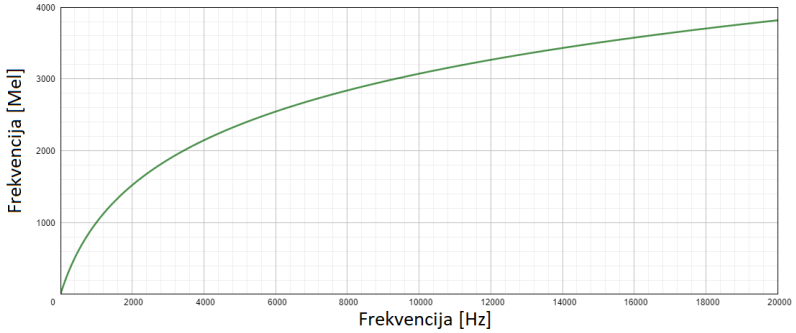


Slika 2.9 Prikaz audio signala ljutitog govora u vremenskoj domeni

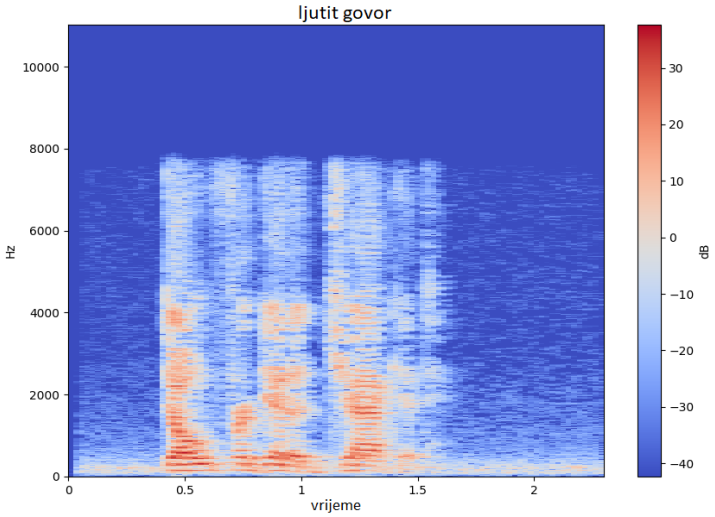
Spektrogramom (Sl. 2.11) je moguće prikazati informacije o audio signalu pomoću vremenske (x) i frekvencijske (y) osi te bojom jačinu, odnosno glasnoću signala. Kako govor nije linearan (frekvencije u govoru se mijenjaju), za dobivanje spektrograma, prema [31], audio signal se uokviruje i nad njim se primjenjuje funkcija prozora (engl. *windowing*) te se nad svakim okvirom provodi Brza Fourierova Transformacija (engl. *Fast Fourier Transform*). Algoritam koji objedinjuje navedene potrebne korake za pretvorbu podataka za prikaz spektrograma naziva se Kratkovremena Fourierova Transformacija (engl. *Short-time Fourier Transformation*). Spektrogram se promatra kao prikaz naslaganih Fourierovih transformacija. Za detaljniji prikaz informacija na spektrogramu, vrijednosti amplituda pretvaraju se u decibele i frekvencijska (y) os se pretvara u logaritamsku skalu umjesto linearne skale. Ljudska bića ne razlikuju frekvencije na linearnoj skali. Primjerice, ljudi primjećuju razliku između 500 i 1000 Hz dok teže primjećuju razliku između 10000 i 10500 Hz iako je apsolutna razlika od 500 Hz jednaka. Radi toga je prihvaćena Mel skala, koja je u srži logaritamska transformacija frekvencije signala. Termin Mel je skraćenica od engleske riječi *melody*. Pretvorba iz Hz skale u Mel skalu izvodi se pomoću izraza

$$m = 1127 \ln\left(1 + \frac{f}{700}\right), \qquad (2\text{-}8)$$

gdje m predstavlja vrijednost u Mel skali, a f vrijednost frekvencije u Hz skali. Na slici 2.10 prikazan je odnos između Mel skale i frekvencije izražene u Hz.



Slika 2.10 Grafički prikaz odnosa Mel skale i frekvencije u Hz



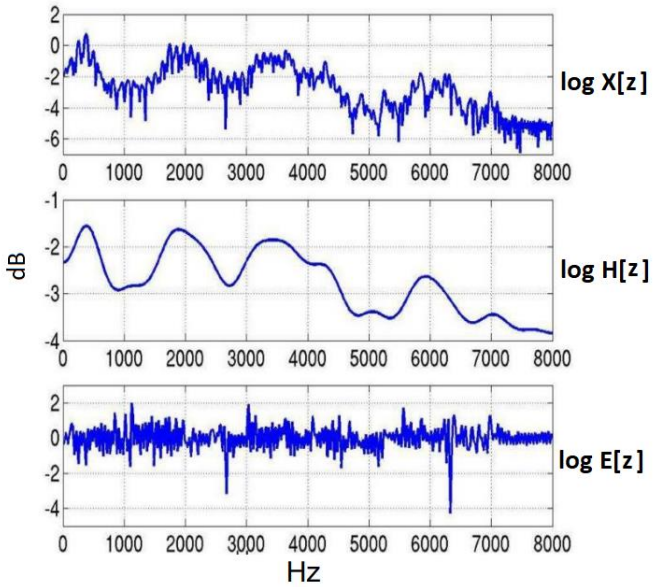
Slika 2.11 Prikaz ljutitog govora spektrogramom

Mel spektrogram predstavlja spektrogram kojemu su vrijednosti frekvencijske (y) osi skalirane prema Mel skali. Glasnoća na Mel spektrogramu je označena bojom te su vrijednosti glasnoće u decibelima [31]. Osim značajki Mel spektrograma, popularne spektralne značajke su MFCC značajke koje su zasnovane na modelu ljudskog sluha. Prema već navedenom, ljudi doživljavaju zvukove po Mel skali koja nije linearna. Prema slici 2.10, frekvencije do 1 kHz se doživljavaju približno linearno dok se nakon frekvencije od 1 kHz doživljavaju po logaritamskoj skali Kako je za ovaj projekt naglasak stavljen na obradu glasovnog zapisa, potrebno je razmatrati proizvodnju glasa na temelju koncepta izvor-filter, gdje je izvor predstavljen zrakom kojeg se oslobađa iz pluća, a filter ljudskim vokalnim traktom koji se sastoji od dijelova poput nosne i usne šupljine, jezika, glasnica i dušnika. Ispuštanjem zraka iz pluća i oblikovanjem vokalnog trakta proizvode se

određeni glasovi. Nadalje, glas se može promatrati kao konvolucija signala izvora e[n] i filtera h[n], prema

$$x[n] = e[n] * h[n]. \quad (2\text{-}9)$$

Za potrebe raspoznavanja emocija iz glasovnog zapisa, najvažnije informacije nalaze se u signalu vokalnog trakta, odnosno filteru. Osnovna frekvencija govornika, koja proizlazi iz signala izvora, ne pridonosi u procesu raspoznavanja emocija iz glasa te je potrebno izdvojiti filter iz glasovnog signala. Primjer dekonvolucije signala tj. izdvajanja signala filtera od signala izvora prikazan je na slici 2.12. Upravo na tim informacijama filtera, koje se produciraju oblikom vokalnog trakta pri izgovoru, temelje se MFCC značajke.

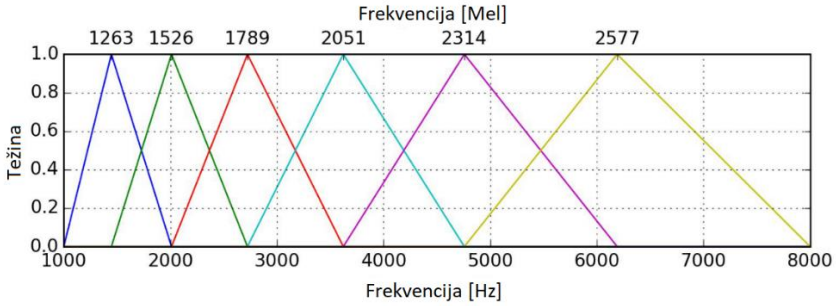


Slika 2.12 Grafički prikaz dekonvolucije signala izvora od filtera [32]

Postupak izdvajanja MFCC značajki, prema [33], započinje transformacijom uokvirenog signala iz vremenske u frekvencijsku domenu, odnosno Fourierovom transformacijom:

$$X[z] = E[z] * H[z] \quad (2\text{-}10)$$

Nakon dobivenog spektra mjerenog u Hz skali, spektar se prilagođava ljudskom sluhu prema Mel skali primjenom skupa filtera po Mel skali (engl. *Mel filter bank*), koji je preklapajućeg trokutastog oblika, kao na slici 2.13. te se amplitude spektra skaliraju po logaritamskoj, odnosno decibel skali.

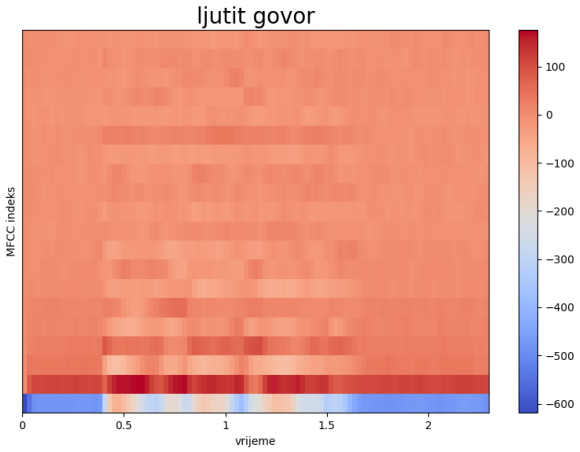


Slika 2.13 Grafički prikaz skupa filtera po Mel skali [32]

U konačnici, nad logaritamskim spektrom, provodi se inverzna diskretna Fourierova transformacija (u praksi najčešće inverzna diskretna kosinusna transformacija) koja pretvara iz frekvencijske domene u pseudo frekvencijsku domenu poznatu kao *quefrency* (anagram engleske riječi *frequency*) domena te kao rezultat daje *cepstrum* (anagram engleske riječi *spectrum*). *Cepstrum* se može promatrati kao spektar spektra koji se matematički može zapisati kao

$$C\big(x(t)\big) = F^{-1}[\log\,(F\big(x(t)\big)], \qquad (2\text{-}11)$$

gdje x(t) predstavlja signal u vremenskoj domeni, F Fourierovu transformaciju. Amplitudne vrijednosti u dobivenom *cepstrum*-u su MFCC značajke. Prema [34], tipično se koristi oko 20 MFCC koeficijenata za potrebe računalnog raspoznavanja emocija i smatra se da najveći nedostatak pri korištenju MFCC značajki predstavlja osjetljivost na šumove zbog ovisnosti o spektralnom obliku. Prema slici 2.14, prikazani su MFCC značajke za glasovni zapis ljutitog govora. Odabrano je 20 MFCC značajki. Svaki stupac predstavlja jedan vremenski okvir, dok svaki redak predstavlja jedan izdvojeni koeficijent.



Slika 2.14 Prikaz MFCC značajki za ljutit govor

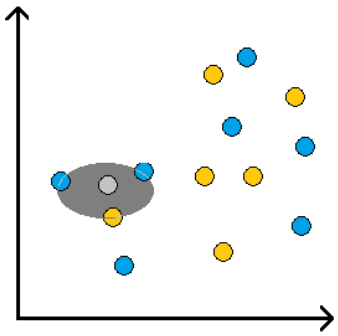
2.4.5. Često korišteni klasifikatori

Klasifikator je algoritam koji na temelju ulaznih podataka predviđa klasu. Pri govornom raspoznavanju emocija, ulazne podatke predstavljaju izdvojene i pripremljene značajke dok izlazni sloj predstavljaju diskretne klase emocija. Unatoč velikom broju klasifikatora, ne postoji klasifikator koji je pogodan za sve oblike problema strojnog učenja. Često korišteni klasifikatori, prema [35, 36], pri govornom raspoznavanju emocija su:

- **K-najbližih susjeda** (engl. *k-nearest neighbour*), prema [35] je lijeni klasifikator koji na temelju podataka za treniranje računa udaljenost točaka u n-dimenzionalnom prostoru, te na temelju oznaka klasa k najbližih susjednih točaka odlučuje kojoj klasi pripada tako što novom uzorku dodjeljuje oznaku većinske klase među susjedima (Sl. 2.15). Udaljenost se najčešće računa pomoću euklidske udaljenosti

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (2\text{-}12)$$

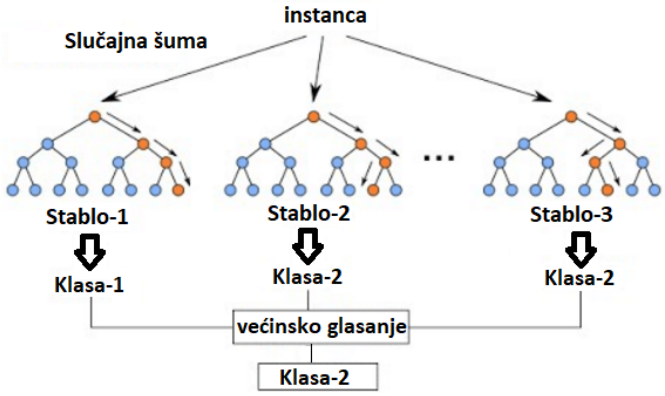
  Kao i za većinu klasifikatora, parametri se odabiru proizvoljno, a za početnu vrijednost parametra k se često preporučuje da bude neparan broj i da iznosi otprilike korijen od ukupnog broja podataka za treniranje.



Slika 2.15 Prikaz glasovanja k-NN klasifikatora gdje je k=3 u dvodimenzionalnom prostoru, izrađeno prema [37]

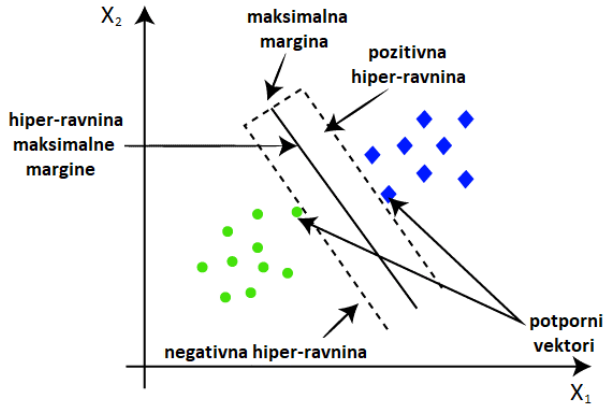
- **Slučajna šuma** (engl. *random forest*), prema [35] je klasifikator koji se sastoji od stabala odluke. Svako stablo se sastoji od čvorova koji predstavljaju if-uvjet, grana koje predstavljaju ishode i listova koji predstavljaju klase. Klasifikacija se odvija tako što

podaci prolaze kroz stabla odluke i svako stablo predviđa klasu te svaki rezultat odlazi na glasovanje, gdje većinski broj glasova određuje klasu (Sl. 2.16).



Slika 2.16 Prikaz klasifikacije pomoću slučajne šume, izrađeno prema [38]

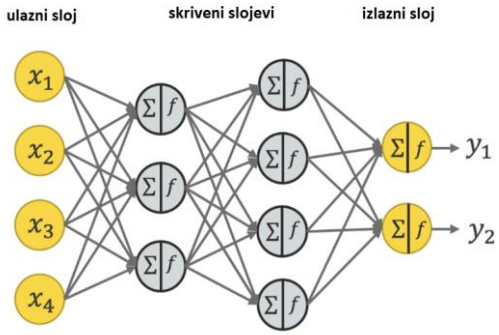
- **Stroj s potpornim vektorima** (engl. *support vector machine*, SVM), prema [35] je klasifikator kojemu je cilj kreirati granicu kako bi razdvojio n-dimenzionalni prostor u klase. SVM odabire podatkovne točke u n-dimenzionalnom prostoru različitih klasa koje su međusobno najbliže kako bi na temelju njih kreirao hiper-ravninu. Klasifikacija se odvija tako što se nova instanca smješta u n-dimenzionalni prostor te se na temelju položaja određuje klasa. Prema slici 2.17, mogu se primijetiti točke u dvodimenzionalnom prostoru koje predstavljaju potporne vektore, nad kojima su se izradile hiper-ravnine koje odvajaju jednu klase od druge.



Slika 2.17 Prikaz klasifikacije pomoću stroja s potpornim vektorima, izrađeno prema [39]

- **Neuronska mreža** (engl. *neural network*), prema [36] je klasifikator koji se sastoji od 3 sloja: ulazni, skriveni i izlazni (Sl. 2.18). Ulazni sloj prima podatke i prosljeđuje ih skrivenom sloju koji izvodi nelinearne transformacije nad primljenim podacima. Izlazni

sloj predstavlja krajnji rezultat problema te čvorovi mogu predstavljati klase i sadržavati vrijednost koja će predstavljati vjerojatnost te klase za unesen ulaz. Čvorovi skrivenog sloja su aktivacijske funkcije koje predaju rezultat sljedećim čvorovima s kojima su povezani.

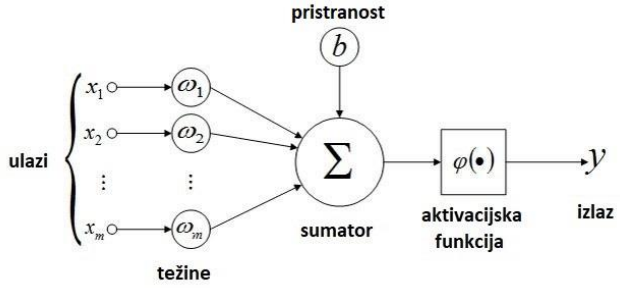


Slika 2.18 Prikaz modela neuronske mreže [40]

Treniranjem neuronske mreže smatra se podešavanje težina (engl. *weights*) veza između čvorova (neurona) koje se odvija algoritmom propagacije pogreške unatrag (engl. *backpropagation algorithm*). Prema slici 2.19, blokovskom shemom prikazan je model neurona koji prima ulazne podatke (vektor x), nad kojima su primijenjene vrijednosti težina (vektor ω) te se sve zbrajaju, nakon čega im se pridodaje vrijednost pristranosti (engl. *bias*) te ulaze u aktivacijsku funkciju koja dalje prosljeđuje rezultat na obradu nekom sljedećem neuronu ili šalje rezultat izlaznom sloju. Matematički izraz za neuron prvog sloja bi glasio

$$y = \varphi(\sum_{i=1}^{n}(x_i w_i) + b)\,, \qquad (2\text{-}13)$$

gdje y predstavlja izlaznu vrijednost iz neurona, x ulaznu vrijednost, w težinu veze i b pristranost neurona.



Slika 2.19 Matematički model neurona [41]

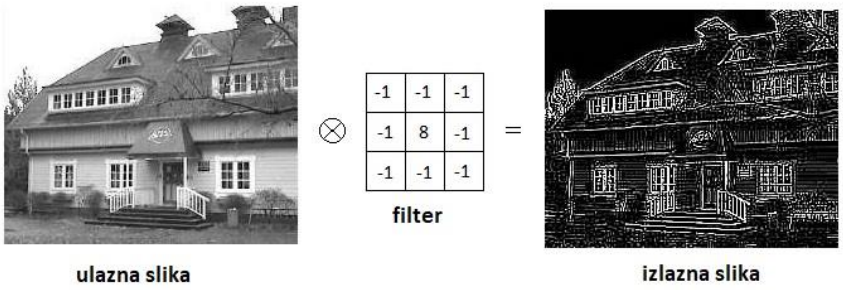
Osim klasičnog strojnog učenja za potrebne raspoznavanja emocija aktualno je duboko učenje. Često se pojmovi poput strojnog učenja, dubokog učenja i umjetne inteligencije koriste kao međusobno zamjenski pojmovi, ustvari oni se promatraju kao podskup skupa. Duboko učenje je podskup skupa neuronskih mreža, neuronske mreže su podskup skupa strojnog učenje i na kraju strojno učenje je podskup skupa umjetne inteligencije. Razliku između klasičnog strojnog učenja neuronskih mreža i dubokog učenja čini veći broj skrivenih slojeva. Veći broj skrivenih slojeva omogućuje dubokoj neuronskoj mreži samostalno izvlačenje potrebnih značajki bez ljudske intervencije. Značajnu razliku pri treniranju mreže se primjećuje u vremenu treniranje gdje duboka mreža zahtijeva više vremena, resursa i znatno veće količine podataka [42].

- **Konvolucijska neuronska mreža**

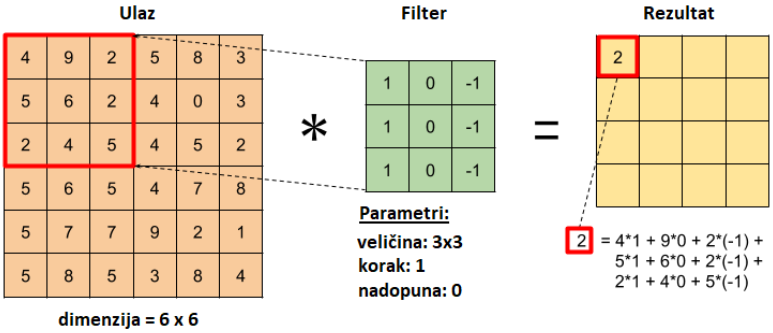
Konvolucijska neuronska mreža (engl. *convolutional neural network (CNN)*) (Sl. 2.25) predstavlja evoluiranu klasičnu neuronsku mrežu koja je specijalizirana za duboko učenje temeljeno na nestrukturiranim podacima kao što su slike, govor, zvuk i slično. Kao što je već rečeno, konvolucijska neuronska mreža razlikuje se od klasičnog modela neuronske mreže po velikom broju skrivenih slojeva. Komponente CNN mreže čine:

- Ulazni sloj predstavlja ulaznu sliku koja može biti u boji (3 kanala) ili monokromatska (1 kanal), primjerice za sliku dimenzije 64x64 piksela u boji, ulazni sloj će biti dimenzije 64x64x3.
- Konvolucijski sloj (po kojemu je i mreža dobila ime) je sloj mreže koji pomoću matematičke operacije konvolucije rezultira novim signalom, odnosno mapom značajki (engl. *feature map*) koja pruža više informacija od izvornog signala. Primjer detekcije rubova fotografije prikazan je na slici 2.20. Konvolucijski sloj se sastoji od filtera, najčešće veličine 3x3, gdje je svaki filter konvoluiran s ulaznom slikom i rezultira mapom značajki. Primjer izračuna jednog koraka jednostavne konvolucije prikazan je na slici 2.21. Parametri filtera unutar konvolucijskog sloja nisu predefinirani već se treniraju prilikom učenja mreže.

  Slaganjem konvolucijskih slojeva prilikom dizajniranja konvolucijske neuronske mreže, kao što je već i rečeno, postiže se izdvajanje značajki. Prvim slojem se izdvajaju značajke niske razine, te svim sljedećim slojevima se izdvajaju značajke viših razina, primjerice od rubova i geometrijskih oblika pa postepeno do konkretnih objekata.



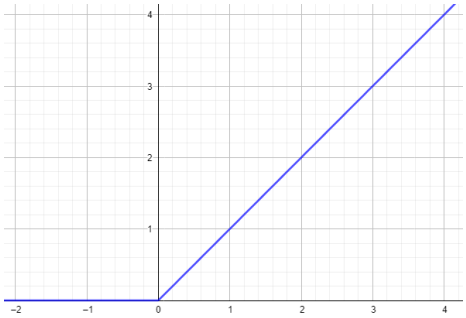
Slika 2.20 Konvolucija filterom za detekciju rubova [43]



Slika 2.21 Konvolucija dvaju signala [44]

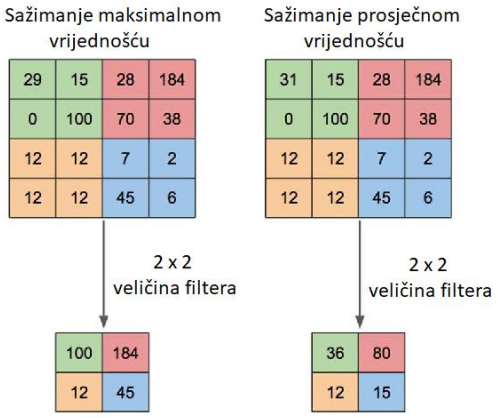
- Aktivacijska funkcija primjenjuje nelinearnu transformaciju nad ulaznim podacima te time omogućava mreži učenje i rješavanje složenih zadataka. Bez aktivacijskih funkcija unutar neuronske mreže, svi slojevi bi se jednako ponašali zbog svojstva linearnosti. Najčešće korištena aktivacijska funkcija u dizajniranju konvolucijske neuronske mreže je ReLu (*Rectified linear unit*) funkcija (Sl. 2.22) koja omogućuje brže i učinkovitije treniranje mreže tako što negativne vrijednosti mapira u 0 i čuva pozitivne vrijednosti te se aktivirane (pozitivne) značajke se prenose u sljedeći sloj. Matematički je zapisana kao

$$f(x) = \max(0, x). \qquad (2\text{-}14)$$



Slika 2.22 Grafički prikaz ReLu funkcije

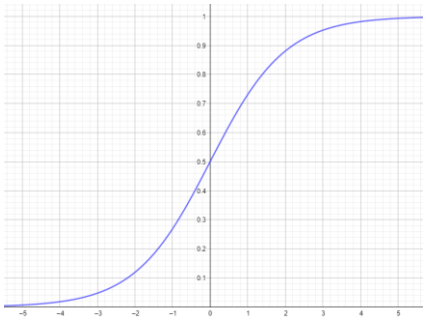
- Sloj sažimanja, prema [45, str. 442], je sloj kojemu je cilj smanjiti ulaznu sliku, odnosno mapu značajki, zbog smanjenja vremena izračuna računskih radnji, korištenja memorije i broja parametara neuronske mreže te time i izbjeći problem pretjerane prilagodljivosti (engl. *overfitting*). Sažimanje se provodi tako da se filterom prolazi nad mapom značajki te se za svaki odjeljak mape odabire vrijednost po nekom od algoritama sažimanja. Osnovni algoritmi sažimanja su sažimanje izborom maksimalnog elementa (engl. *max-pooling*) i sažimanje prosječnom vrijednošću (engl. *average-pooling*) (Sl. 2.23). Hiper-parametri ovog sloja su pomak i veličina filtera, koji obično budu 2 i veličine 3x3.



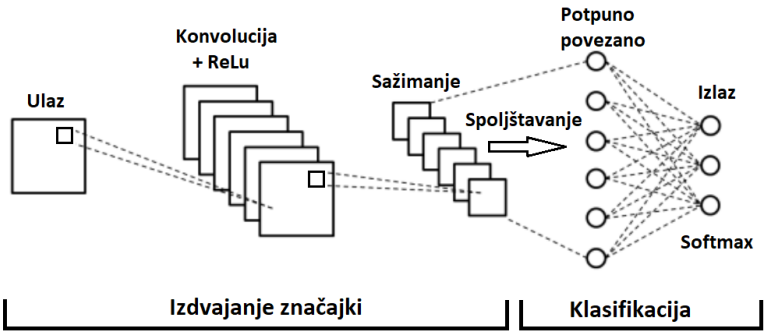
Slika 2.23 Usporedba sažimanja maksimalnom i prosječnom vrijednošću [46]

- Potpuno povezani sloj, kao i kod klasične neuronske mreže, sadrži neurone koje su povezani sa svim neuronima iz prethodnog sloja. Ulaz ovog sloja je posljednji sloj sažimanja ili konvolucijski sloj koji je spljošten u jednodimenzionalni vektor dok izlaz predstavlja vektor s N klasa. Često poslije potpuno povezanog sloja slijedi sloj koji koristi *softmax* aktivacijsku funkciju (Sl. 2.24) koja pruža vjerojatnost za svaku klasu. Matematički zapis *softmax* funkcije glasi

$$f(x) = \frac{1}{1 + e^{-x}}. \qquad (2\text{-}15)$$



Slika 2.24 Grafički prikaz *softmax* funkcije



Slika 2.25 Prikaz modela konvolucijske neuronske mreže, izrađeno prema [47]

2.5. Mogućnosti raspoznavanje emocija na Android platformi

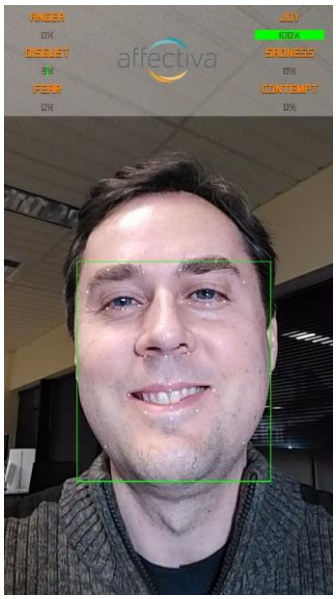
Jedan način puštanja modela strojnog učenja u produkciju bio bi putem web servisa, koji bi predstavljao REST API uslugu. Modelu se dostavljaju podaci putem POST metode i u konačnici web servis pruža rezultat u JSON (JavaScript Object Notation) formatu. Za izrađeni model u Python-u, pomoću okvira (engl. *framework*) Flask, model se prenosi na odabrani web servis te je spreman za klasifikaciju. Na Android uređaju je potrebno pripremiti instancu emocije s odgovarajućim značajkama. I kako bi Android uređaj slao i primio podatke s vanjskog servisa (u ovom slučaju REST API) koristi se biblioteka Retrofit za HTTP komunikaciju. Drugi način korištenja modela strojnog učenja na Android platformi je ugraditi gotov model. Izrađeni model se pretvara u TFLite datoteku pomoću TFLite pretvarača (engl. *converter*) i uvozi se u Android projekt, također je se uvozi TensorFlowLite paket u ovisnosti (engl. *dependencies*). Za proces klasifikacije, kao i kod prvog načina, potrebno je pripremiti instancu emocije te kreirati objekt prevoditelja (engl. *interpreter*) koji prima instancu emocije i vraća predviđenu klasu emocije [48].

2.5.1. Postojeća rješenja

Vokaturi (Sl. 2.26) je biblioteka za raspoznavanje emocija na temelju snimljenog audiozapisa. Može raspoznati 5 emocija: sreću, tugu, ljutnju, strah i neutralnost. Biblioteka je dostupna za iOS, Windows, MacOS i Android. Implementacija na Android platformi je vrlo jednostavna, temelji se na instanciranju *singleton* objekta Vokaturi klase koji omogućuje snimanje glasa i kao krajnji rezultat pruža objekt koji sadrži vjerojatnosti za svih 5 emocija [49].

Slika 2.26 Prikaz demo Vokaturi aplikacije [49]

AffdexMe (Sl. 2.27) je aplikacija koja analizira i prikazuje trenutnu emociju na temelju izraza lica u stvarnom vremenu, Dostupna je za Android i iOS uređaje. Aplikacija omogućava još i praćenje željene emocije, prebacivanje između prednje i stražnje kamere, pohranjivanje slike u galeriju, prikaz točaka lica na temelju kojih se izračunavaju vjerojatnosti emocija, praćenje više lica odjednom [50].



Slika 2.27 Prikaz korištenja AffdexMe aplikacije [50]

Emotimeter - Emotion detector (Sl. 2.28) je također aplikacija za raspoznavanje emocija iz izraza lica. Pruža raspoznavanje emocija u stvarnom vremenu pomoću kamere, fotografiranje ili

snimanje videozapisa tijekom sesije raspoznavanja emocija, analiza fotografija i videozapisa iz galerija u svrhu raspoznavanja emocija svih osoba u fotografiji ili videozapisu [51].

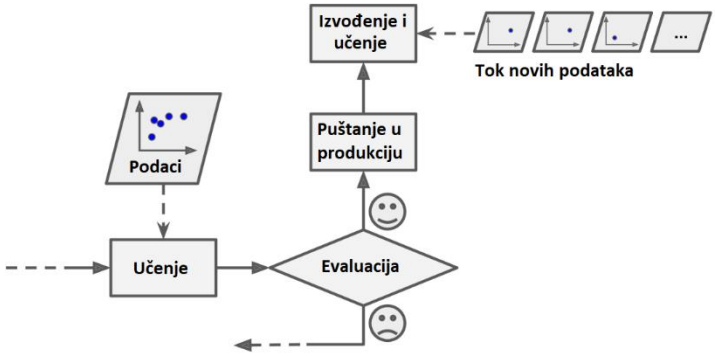
Slika 2.28 Prikaz korištenja Emotimeter aplikacije [51]

Group Emotion Recognition - Detect Face Expression (Sl. 2.29) je programsko rješenje implementirano na Android platformi koja pomoću dva modula (odozgo prema dolje i odozdo prema gore) raspoznaje emocije osoba na fotografijama. Modul odozdo prema gore koristi konvolucijsku neuronsku mrežu za detekciju emocije iz izraza lica na fotografiji. Modul odozgo prema dolje koristi Bayesovu mrežu (engl. *Bayesian Network*) koja detektira oznake (engl. *labels*) s fotografije. Oba modula rade istovremeno [52].

Slika 2.29 Prikaz korištenja Group Emotion Recognition aplikacije [52]

2.5.2. Prednosti i nedostatci online i offline pristupa

Sustave strojnog učenja, prema [45, str. 15], moguće je podijeliti na dvije skupine, skupinu koja uči inkrementalno iz tokova podataka ili skupinu koja je naučena na prikupljenoj gomili podataka. Treniranje gomile podataka (engl. *batch learning*) ili poznato još kao i izvanmrežno učenje (engl. *offline learning*) predstavlja oblik treniranja modela strojnog učenja tako što se uči na gomili podataka i nema mogućnost inkrementalnog učenja iz nekakvog toka podataka (engl. *stream of data*). Nedostatak ovog oblika treniranja leži u tome što koristi puno računalnih resursa i zahtijeva mnogo vremena. U slučaju ažuriranja modela strojnog učenja novim podacima potrebno je nanovo istrenirati model s novim i sa starim podacima, postupak je moguće automatizirati. Ako je model postavljen kao usluga na poslužitelju, nakon što je novi model nanovo istreniran nad novim i starim podacima, potrebno je zaustaviti trenutnu uslugu i zamijeniti stari model novim modelom. U slučaju postavljenog modela unutar uređaja, postupak treniranja novog modela je isti te nakon zamjene starog modela novim je potrebno objaviti novu verziju aplikacije kako bi uređaji ažurirali aplikaciju novom verzijom. Ovaj oblik treniranja je pogodan za sustave koji ne zahtijevaju česte prilagodbe. Online učenje (engl. *online learning*) je oblik treniranja gdje model strojnog učenja inkrementalno napreduje i gdje je svako novo učenje brzo i ne zahtijeva puno resursa (Sl. 2.30). Ovakav oblik učenja poželjan je za sustave strojnog učenja koji primaju nove podatke u neprekidnom toku i koji se trebaju brzo prilagoditi. Primjer koji zahtjeva ovakvo učenje bio bi sustav koji predviđa cijene na tržištu. Bitna osobina online učenja je stopa učenja (engl. *learning rate*), odnosno brzina kojom se model strojnog učenja prilagođava novim podacima. Prevelika stopa učenja može uzrokovati da model brzo „zaboravlja“ naučeno, dok premala stopa učenja će biti vrlo slična treniranju gomile podataka. Problemi mogu nastati ako novi podaci koji pristižu za učenje su ustvari loši podaci i time se uzrokuju loše performanse sustava.



Slika 2.30 Shematski prikaz online učenja, izrađeno prema [45, str. 16]

Odabrana varijanta za ovaj projekt je treniranje gomile, iz razloga što je upotreba u produkciji jednostavnija od online varijante te zbog jednostavnije implementacije. Pod pojmom jednostavne upotrebe u produkciji smatra se jednostavno rukovanje performansama klasifikacijskog modela za vrijeme produkcije. Također, razlog odabira treniranje gomile je zbog ograničenosti besplatnih poslužitelja u smislu računalne snage za obradu neprekidnog toka podataka koju zahtijeva online varijanta.

3. ALAT ZA RASPOZNAVANJE EMOCIJA I SNIMLJENOG GOVORA *KNOW YOURSELF*

Know yourself je aplikacija za Android mobilne uređaje koja služi korisniku za raspoznavanje emocija (sreće, tuge, gađenja, straha, ljutnje i neutralnog osjećaja). Snimanjem glasovnog zapisa te korištenjem vanjskih servisa korisniku se prikazuju odgovarajuće informacije o raspoznatoj emociji. Na taj način korisnik se informira te nauči vladati svojim vlastitim emocijama i podiže svoju razinu emocionalne inteligencije. Osim vizualnog sadržaja (fotografije i teksta), korisnik ima mogućnost reproducirati audiozapis odgovarajuće *Solfeggio* frekvencije. Prema [53], *Solfeggio* frekvencije su dijelovi glazbene ljestvice koja je korištena u starim gregorijanskim napjevima te one utječu na emocionalno stanje korisnika. Pojam *solfeggio* predstavlja dodjeljivanje sloga notama glazbene ljestvice. Originalno postoji 6 frekvencija, odnosno tonova, koji su preuzeti iz srednjovjekovne himne svetom Ivanu Krstitelju. Naknadno su dodane još 3 frekvencije. Tako je za potrebe ove aplikacije, moguće je emocije povezati s jednom od 9 *Solfeggio* frekvencija. Također, korisnik može pratiti svoje emocionalno stanje kroz zadnjih 7 dana, vizualizacijom lokalno pohranjenih raspoznatih emocija po danu. Za jasni prikaz funkcionalnosti Know yourself aplikacije dana je tablica 3.1 koja sadrži potrebne zahtjeve. Također, navedeni su slučajevi korištenja aplikacije koji su detaljno opisani u sljedećem potpoglavlju.

ID	Opis	UC
0	Izrada REST API servisa za raspoznavanje emocija	-
1	Aplikacija prikazuje početni meni	-
2	Korisnik snima glasovni zapis	UC0
3	Aplikacija pohranjuje glasovni zapis u *.wav* formatu	-
4	Aplikacija koristi REST API za raspoznavanje emocija	-
5	Validacija korisnika za raspoznatu emociju	UC1
6	Aplikacija koristi vanjske servise za prikaz potrebnog sadržaja	-
7	Aplikacija prikazuje poruke o pogreškama koristeći Android *Toast* poruke	-
8	Aplikacija pohranjuje raspoznate emocije lokalno	-
9	Korisnik može reproducirati dobivenu *Solfeggio* frekvenciju i upravljati putem notifikacije	UC2
10	Korisnik može dohvatiti vizualni prikaz raspoznatih emocija u zadnjih 7 dana	UC3

Tablica 3.1 Tablica zahtjeva programskog rješenja

3.1. Slučajevi korištenja

Aplikacija je zamišljena s funkcionalnostima kao što su snimanje glasovnog zapisa, raspoznavanje emocije iz snimke glasovnog zapisa, validacija raspoznate emocije te prikaz informacija o raspoznatoj emociji i vizualizacija raspoznatih emocija u posljednjih 7 dana. Na uzoru ovih slučajeva korištenja u budućem poglavlju će biti detaljan prikaz dijagrama toka aplikacije.

ID	UC0
Ime	Snimanje glasovnog zapisa
Opis	Korisnik snima glasovni zapis za daljnju obradu i dohvaćanje raspoznate emocije
Preduvjet	Korisnik je dopustio potrebna dopuštenja za snimanje glasovnih zapisa
Glavni scenarij	1. Korisnik drži tipku za snimanje glasovnog zapisa 2. Korisnik odgovara na postavljeno pitanje 3. Korisnik pušta tipku za snimanje glasovnog zapisa
Alternativni scenarij	1. Korisnik nije glasovno odgovorio na pitanje 1.1. Prikazuje se *Toast* poruka 1.2. Povratak na 1. korak glavnog scenarija 2. Korisnik prerano pušta tipku za snimanje glasovnog zapisa 2.1. Prikazuje se *Toast* poruka 2.2. Povratak na 1. korak glavnog scenarija

ID	UC1
Ime	Validiranje raspoznate emocije
Opis	Korisnik odgovara na pitanje o točnosti raspoznavanja po njegovoj procjeni
Preduvjet	Korisnik je snimio glasovni zapis i složio se za korak raspoznavanja emocije te ima pristup Internetu
Glavni scenarij	1. Korisnik klikom na gumb potvrde slaže se s raspoznatom emocijom
Alternativni scenarij	1. Korisnik se ne slaže s raspoznatom emocijom 1.1. Korisnik odabire gumb za promjenu raspoznate emocije 1.2. Korisnik odabire 1 od 6 mogućih emocija

ID	UC2
Ime	Reproduciranje audiozapisa (*Solfeggio* frekvenciju)
Opis	Korisnik upravlja reprodukcijom audiozapisa korisničkim sučeljem unutar aplikacije ili putem notifikacije Android sustava
Preduvjet	Korisnik je snimio glasovni zapis i validirao raspoznatu emociju
Glavni scenarij	1. Korisnik povlači fragment od dolje prema gore za prikaz fragmenta

	2. Korisnik za reprodukciju audiozapisa koristi gumb za pokretanje 3. Korisnik za zaustavljanje audiozapisa koristi stop gumb
Alternativni scenarij	1. Korisnik zaustavlja audiozapis putem notifikacije Android sustava 1.1. Pritiskom na stop gumb na notifikaciji, audiozapis se zaustavlja i notifikacija briše

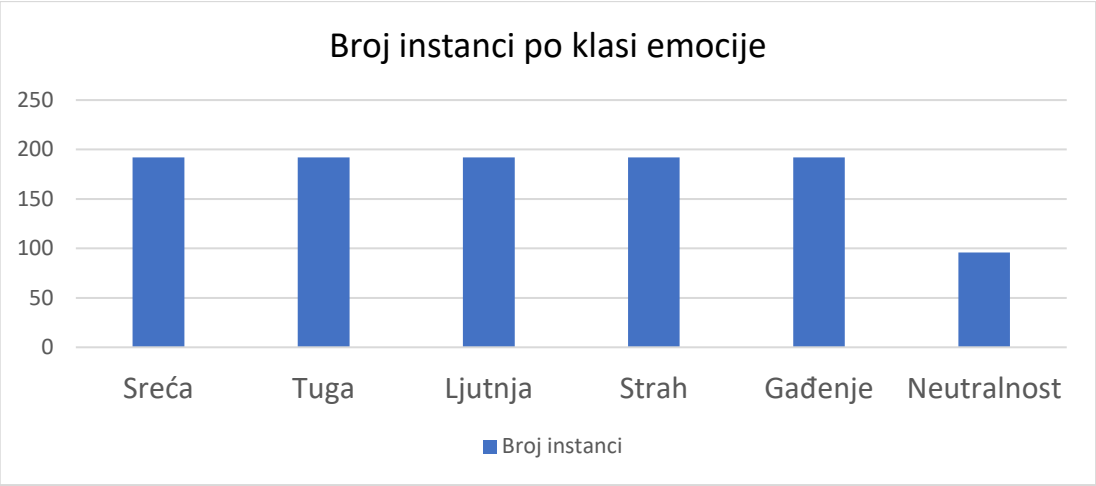
ID	UC3
Ime	Vizualizacija pohranjenih raspoznatih emocija
Opis	Korisnik odabire u glavnom meniju gumb za vizualizaciju podataka te mu se prikazuju emocije tijekom zadnjih 7 dana
Preduvjet	Postoje pohranjene emocije kroz zadnjih 7 dana
Glavni scenarij	1. Korisniku se prikazuju raspoznate emocije kroz zadnjih 7 dana
Alternativni scenarij	1. Ne postoje pohranjene emocije kroz zadnjih 7 dana 1.1. Za svaki dan koji nije pohranjena raspoznata emocija prikazuje se odgovarajuća poruka

3.2. Eksperiment za izgradnju modela

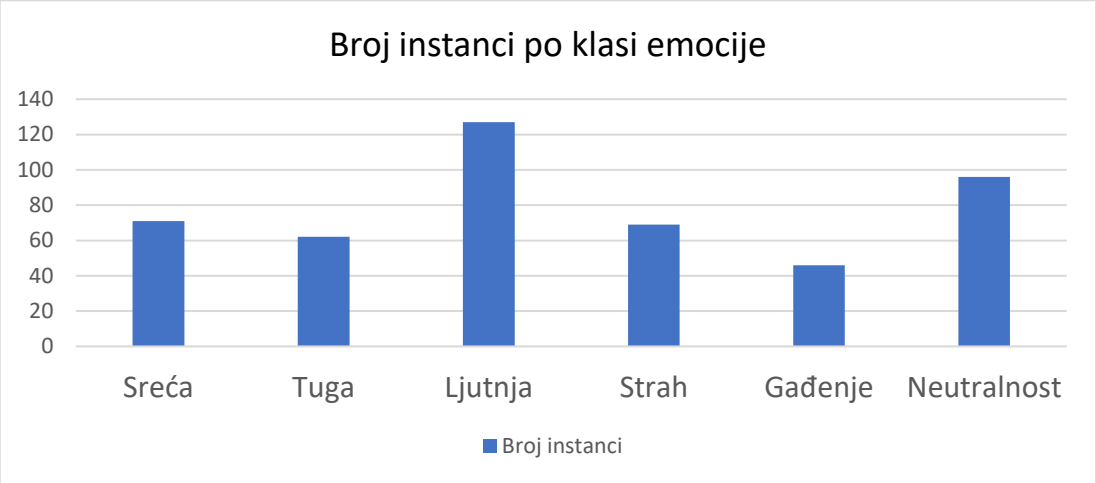
Buduća potpoglavlja donose postavljanje i pripremu eksperimenta, analizu triju različitih modela učenih nad dvije različite baze i postupak postavljanja klasifikacijskog modela u aplikaciju. Cilj eksperimenta je pronaći model s najboljim rezultatima koji će biti korišten u implementaciji Android aplikacije. Istrenirani model će klasificirati snimke govora u jednu od 6 klasa, odnosno emocija sreće, tuge, ljutnje, gađenja, straha i neutralne emocije.

3.2.1. Postavljanje eksperimenta

Eksperiment će se izvesti nad *RAVDESS* i *Emo-DB* bazom podataka te su za ovaj rad odabrane emocije sreće, tuge, ljutnje, straha, gađenja i neutralnog osjećaja. Iz *RAVDESS* baze podataka izabrano je 1056 instanci, dok za *Emo-DB* bazu 454 instanci. Raspodjela instanci po emocijama za *RAVDESS* bazu prikazana je na grafikonu 3.1, a za *Emo-DB* bazu na grafikonu 3.2. Baze su odabrane zbog manjeg šuma i veće jasnoće raspoznavanja emocija sluhom za razliku od drugih potencijalnih baza za ovaj projekt. Snimljeni audiozapisi *RAVDESS* baze su prilično tihi te će se pomoću Pydub [54] biblioteke pojačati za 15dB.



Grafikon 3.1 Prikaz raspodijele instanci emocija za *RAVDESS* bazu podataka



Grafikon 3.2 Prikaz raspodijele instanci emocija za *Emo-DB* bazu podataka

Kako je već spomenuto u poglavlju 2.4.4., spektralne značajke prigodne su za višeklasni problem raspoznavanja emocija. Preliminarnim ispitivanjem utvrdilo se da značajke Mel spektrograma ne postižu dobre rezultate te time će klasifikatori biti učeni na temelju MFCC značajki. Osim Mel značajki, preliminarnim ispitivanjem je i k-najbližih susjeda klasifikator pokazao loše rezultate te stoga njegovi rezultati nisu uključeni u usporedbu već je isključen iz daljnjeg razmatranja. Iz baze podataka izdvojiti će se prvih 20 MFCC značajki. Nadalje će se trenirati klasifikatori klasičnog strojnog učenja: slučajna šuma i stroj s potpornim vektorima nad srednjim vrijednostima MFCC značajki. Osim navedenih klasifikatora, trenirat će se i klasifikator dubokog učenja, konvolucijska neuronska mreža koja će biti trenirana nad cijelom slikom MFCC značajki. Oblik MFCC značajki pri treniranju SVM i RF klasifikatora je jednodimenzionalno polje od 20 elemenata, dok je ulazni

oblik (engl. *input shape*) za CNN klasifikator trodimenzionalan i sadrži dimenzije širine, visine i dubine slike, u ovom eksperimentu je oblika (20, 259, 1). Važno je napomenuti kako ulazna slika CNN klasifikatora mora biti fiksna te su zbog toga su svi uzorci ograničeni na duljinu od 3 sekunde. Uzorci koji su dulji od 3 sekunde su skraćeni, a uzorci kraći od 3 sekunde su produljeni tehnikom ispunjavanja signala s desna (engl. *right padding*) što rezultira produljivanjem audiozapisa tišinom. Visina slike je određena brojem vremenskih okvira koji je dobiven prilikom izdvajanja MFCC značajki pomoću Librosa biblioteke. Širina slike je određena brojem MFCC značajki te dubina slike iznosi 1 zbog toga što je slika monokromatska. Zbog manjeg broja uzoraka za učenje duboke neuronske mreže, provest će se povećanje podataka (engl. *data augmentation*). Skaliranje visine tona, dodavanje bijelog šuma i tehnika slučajnog naglašavanja su korištene tehnike povećanja audio uzoraka. Tehnika dodavanja bijelog šuma predstavlja dodavanje nasumičnih uzoraka (raspoređeni u jednakim intervalima) ulaznom signalu, gdje su vrijednosti šuma unutar intervala [-1, 1]. Skaliranje visine tona je tehnika koja mijenja visinu tona audio signala bez utjecaja na brzinu izvođenja tonova. Tehnika slučajnog naglašavanja predstavlja nasumično pojačavanje signala u pravilno raspoređenim intervalima. Svakom klasifikatoru su prilagođeni hiper-parametri u skladu s odabranim značajkama i bazom nad kojom će se testirati. Odabir hiper-parametara klasifikatora klasičnog strojnog učenja je proveden metodom nasumičnog pretraživanja (engl. *randomized search*) iz Python biblioteke Scikit-learn. Dobiveni hiper-parametri prikazani tablicama 3.2, 3.3.

Hiper-parametri	EmoDB	RAVDESS
Broj stabala	150	100
Metoda odabira ukupnog broja značajki	Automatski (sve značajke)	Korijenovanje broja značajki
Maksimalna dubina	100	70
Minimalan broj uzoraka za dijeljenje unutarnjeg čvora	2	2
Minimalan broj uzoraka u čvoru	1	1

Tablica 3.2 Hiper-parametri RF klasifikatora

Hiper-parametri	EmoDB	RAVDESS
Jezgra	Radijalna osnovna funkcija	Radijalna osnovna funkcija
C (parametar regularizacije)	10000	1000
Gamma (doseg utjecaja jednog uzorka)	$1*10^{-5}$	$1*10^{-4}$

Tablica 3.3 Hiper-parametri SVM klasifikatora

Za CNN klasifikator hiper-parametri (Tablica 3.4) su dobiveni ručnim isprobavanjem zbog sporog izvođenja eksperimenta i slabijih performansi računala. Arhitektura CNN klasifikatora treniranog nad EmoDB bazom sastoji se od:

- Konvolucijskog 2D sloja – 32 filtera, 3x3 veličina filtera i ReLu aktivacijske funkcije
- Sloja sažimanja – 3x3 veličine veličina filtera, pomak 2x2 i nadopuna (engl. *padding*) nulama
- Normalizacije serije (engl. *batch normalization*)
- Konvolucijskog 2D sloja – 64 filtera, 3x3 veličina filtera i ReLu aktivacijske funkcije
- Sloja sažimanja – 3x3 veličine veličina filtera, pomak 2x2 i nadopuna nulama
- Normalizacije serije
- Sloja spoljštavanja
- Zbijenog (engl. *dense*) sloja – 128 neurona i ReLu aktivacijska funkcija
- Zbijenog sloja – 6 neurona i Softmax aktivacijska funkcija

te se arhitektura CNN klasifikatora treniranog nad RAVDESS bazom sastoji od:

- Konvolucijskog 2D sloja – 64 filtera, 3x3 veličina filtera i ReLu aktivacijske funkcije
- Sloja sažimanja – 3x3 veličine veličina filtera, pomak 2x2 i nadopuna nulama
- Normalizacije serije
- Sloja isključivanja neurona (engl. *dropout layer*) – stopa isključivanja 0.1
- Konvolucijskog 2D sloja – 64 filtera, 3x3 veličina filtera i ReLu aktivacijske funkcije
- Sloja sažimanja – 3x3 veličine veličina filtera, pomak 2x2 i nadopuna nulama
- Normalizacije serije
- Sloja isključivanja neurona – stopa isključivanja 0.1
- Sloja spoljštavanja
- Zbijenog sloja – 128 neurona i ReLu aktivacijska funkcija
- Sloja isključivanja neurona – stopa isključivanja 0.3
- Zbijenog sloja – 6 neurona i Softmax aktivacijska funkcija.

Hiper-parametri	EmoDB	RAVDESS
Optimizacijski algoritmi	Adam	Adam
Stopa učenja	$5*10^{-5}$	$1*10^{-3}$
Veličina serije	32	64
Broj epoha	15	20

Tablica 3.4 Hiper-parametri CNN klasifikatora

U konačnici, analizom će se odabrati najpogodniji istrenirani model za daljnju primjenu, odnosno model koji će se implementirati u Android aplikaciju za potrebe raspoznavanja emocija. Analiza performansi će se provesti ispitivanjem točnosti, preciznosti, odziva i F1-mjere.

3.2.2. Analiza eksperimenta

Za svaki model, kako bi se izvukle prosječne vrijednosti rezultata, odrađeno je testiranje kroz 20 iteracija. Za klasifikatore klasičnog strojnog učenja, baze su dijeljene na testni i trening skup te je nad trening skupom odrađena k-struka (engl. *k-fold*) unakrsna validacija gdje k iznosi 5. Za CNN klasifikator, također u svakoj klasifikaciji odrađena je podjela podataka ali na 3 skupa, trening, testni i validacijski skup. Udio testnog skupa od ukupnog skupa iznosi 20%, te se od ostalih 80% naknadno dijelilo 30% na validacijski skup i ostatak za trening skup, što u konačnici znači da testni skup iznosi 20%, validacijski skup iznosi 24% i trening skup iznosi 56% od ukupnog broja uzoraka u bazi. Trening skup je korišten za treniranje modela, validacijski skup za optimizaciju hiper-parametara modela i testni skup za objektivnu procjenu performansi naučenog modela. Hiper-parametri klasifikatora su posebno prilagođeni prema vrsti značajki i prema bazi. Mjera kvalitete koja je imala važnu ulogu pri pronalasku odgovarajućih hiper-parametara je logaritamski gubitak (engl. *log loss*).

Prema rezultatima iz tablica 3.5 i 3.6, kao najbolji model se pokazao SVM klasifikator koji je naučen prema EmoDB bazi podataka. Vrijednosti prosječnih preciznosti, odziva i F1-mjere su računate na makro razini umjesto na mikro razini zbog toga što u EmoDB bazi podataka ne postoji ravnoteža između broja instanci po klasi. Prema tablici 3.6, uočljivo je da se F1-mjere značajno razlikuju za pojedine emocije. Klasifikatori trenirani nad EmoDB bazom podataka lošije prepoznaju emocije gađenja, straha i sreće te se CNN klasifikator ističe pri značajno lošijem rezultatu prepoznavanja emocije sreće. RF i SVM klasifikatori trenirani nad RAVDESS bazom imaju relativno slične rezultate, dok CNN treniran nad RAVDESS bazom se razlikuje u tome što lošije prepoznaje neutralnu emociju.

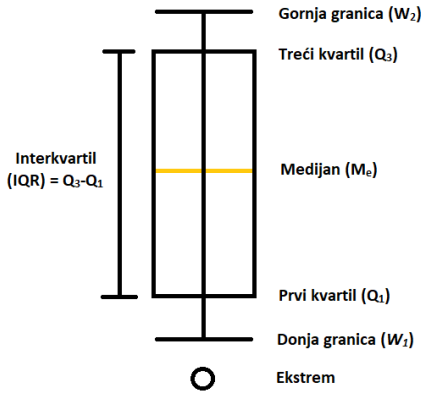
Model	**Baza podataka**	**Preciznost** sr.vr. ± std.	**Odziv** sr.vr. ± std.	**F1-mjera** sr.vr. ± std.	**Točnost** sr.vr. ± std.
RF	RAVDESS	0.649 ± 0.03	0.638 ± 0.03	0.639 ± 0.03	0.64 ± 0.03
RF	EmoDB	0.793 ± 0.05	0.739 ± 0.04	0.744 ± 0.05	0.772 ± 0.03
SVM	RAVDESS	0.637 ± 0.03	0.642 ± 0.03	0.635 ±0.03	0.64 ± 0.03
SVM	**EmoDB**	**0.784 ± 0.03**	**0.773 ± 0.03**	**0.771 ± 0.03**	**0.792 ± 0.02**
CNN	RAVDESS	0.68 ± 0.03	0.6 ± 0.04	0.603 ± 0.04	0.622 ± 0.03
CNN	EmoDB	0.725 ± 0.06	0.695 ± 0.06	0.69 ± 0.06	0.714 ± 0.06

Tablica 3.5 Analiza performansi modela

Model	Baza podataka	Ljutnja	Gađenje	Strah	Sreća	Neutralno	Tuga
RF	RAVDESS	0.705	0.661	0.573	0.633	0.638	0.62
RF	EmoDB	0.815	0.626	0.681	0.579	0.868	0.898
SVM	RAVDESS	0.741	0.632	0.63	0.62	0.61	0.58
SVM	**EmoDB**	**0.853**	**0.661**	**0.695**	**0.618**	**0.871**	**0.926**
CNN	RAVDESS	0.675	0.698	0.636	0.607	0.468	0.538
CNN	EmoDB	0.768	0.669	0.622	0.425	0.789	0.868

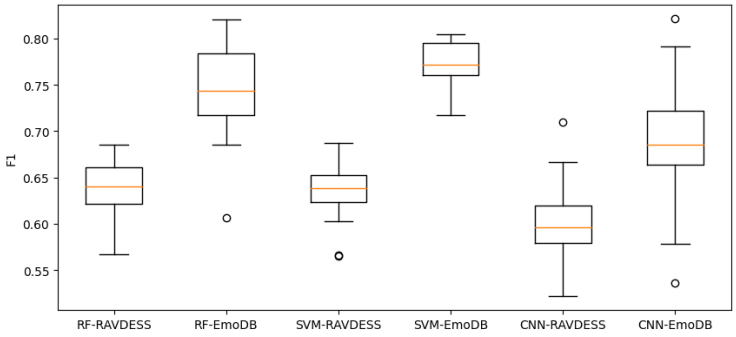
Tablica 3.6 Analiza performansi modela po srednjoj vrijednosti F1-mjere

Analiza je također prikazana kutijastim dijagramom (engl. *box and whisker plots*). Elemente kutijastog dijagrama prikazani su na slici 3.1. Gornja granica ili maksimum predstavlja najveću vrijednost isključujući ekstremne vrijednosti, treći kvartil ili gornji kvartil predstavlja vrijednost ispod koje se nalazi 75% od svih vrijednosti, medijan predstavlja vrijednost koja se nalazi točno u središtu vrijednosti. Sukladno trećem kvartilu, prvi kvartil ili donji kvartil predstavlja vrijednost ispod koje se nalazi 25% od svih vrijednosti i donja granica ili minimum vrijednost koja predstavlja najmanju vrijednost isključujući ekstremne vrijednosti. Kružićem ispod ili iznad interkvartila predstavljena je ekstremna vrijednost koja se značajno razlikuje od raspona većine podataka.



Slika 3.1 Prikaz primjera kutijastog dijagrama

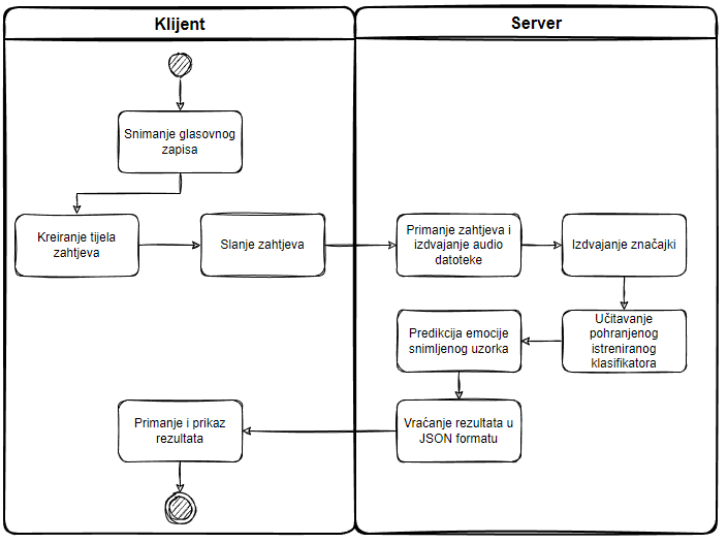
Na slici 3.2, kutijastim dijagramom prikazana je vizualizacija F1-mjere pojedinog klasifikatora nad bazama podataka. Moguće je primijetiti kako su svi klasifikatori ostvarili bolje performanse za EmoDB bazu. Najveću stabilnost pokazao je SVM klasifikator treniran nad EmoDB bazom, dok najmanju stabilnost, odnosno širok raspon F1-mjera pokazao je CNN klasifikator treniran nad EmoDB bazom. Također, stabilnost u smislu F1-mjere pokazuje i SVM klasifikator treniran nad RAVDESS bazom, ali postoji i jedna vrijednost F1-mjere koja je izuzetak, odnosno vrijednost koja ekstremno odstupa od ostalih je prikazana kružićem izvan pravokutnika.



Slika 3.2 Vizualni prikaz performansi modela

3.2.3. Postavljanje modela strojnog učenja u aplikaciju

Za potrebe praktičnog raspoznavanja emocija u daljnjem radu, iz analize eksperimenta odabran je SVM klasifikator treniran nad EmoDB bazom podataka. Navedeni klasifikator je odabran zbog toga što je postigao bolju točnost, preciznost, F1-mjeru i odziv od drugih klasifikatora. Istrenirani klasifikator se pokazao lošijim pri raspoznavanju emocija gađenja, straha i sreće te se zbog toga u aplikaciji zahtijeva validacija raspoznate emocije od korisnika u cilju boljeg korisničkog iskustva. Radi jednostavnosti korištenja gotovog modela, izrađen je REST API servis koji je podignut na *PythonAnywhere* platformi. REST API izrađen je pomoću Python okvira Flask koja omogućava jednostavan i brz razvoj web aplikacija. Kako bi se raspoznala emocija iz .wav formata, POST metodom se šalje audiozapis ključnom riječi „*audioFile*“. Potom se audiozapis učitava i iz njega izvlače značajke pomoću Python biblioteke Librosa te se odrađuje predviđanje istreniranim modelom i u konačnici servis vraća rezultat u JSON formatu. Tako prema slici 3.3, prikazan je shematski prikaz procedure raspoznavanja emocije iz snimke govora.

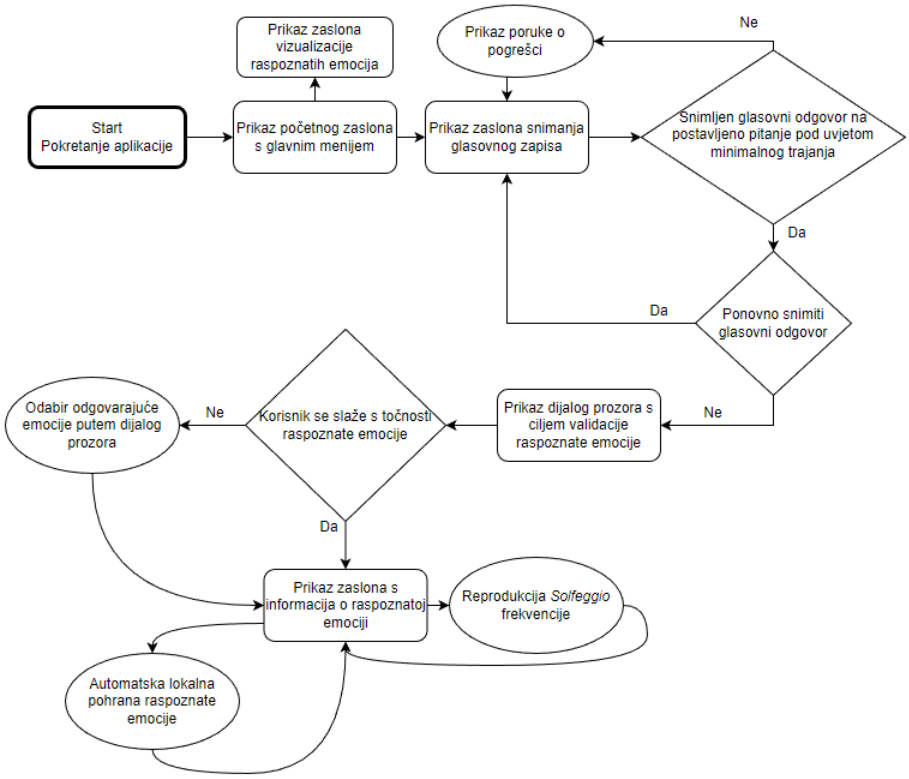


Slika 3.3 Dijagram aktivnosti klijent-poslužitelj komunikacije

Istrenirani model je postavljen kao web usluga umjesto ugrađenog modula unutar aplikacije zbog jednostavnijeg izdvajanja značajki putem Python biblioteka i zbog manje ovisnosti aplikacije o klasifikacijskom modelu. U slučaju ažuriranja klasifikacijskog modela, aplikacija ostaje nepromijenjena i nije usko vezana uz proceduru koja se odvija na poslužiteljskoj strani. Android aplikacija koja zahtijeva odgovor od ovog API servisa, koristi Retrofit biblioteku za HTTP mrežnu komunikaciju.

3.3. Prikaz rada aplikacije

Aplikacija je izrađena prema opisanim specifikacijama i zahtjevima, a pregled rada aplikacije na visokoj razini dan je dijagramom toka. Također su dane detaljne upute za korištenje aplikacije. Dijagram toka izrađen je pomoću zaobljenog pravokutnika koji predstavlja proceduru, elementa odluke i elipse koja predstavlja aktivnost. Upute za korištenje aplikacije su detaljno raspisane po uzoru na dijagram toka (Sl. 3.4).



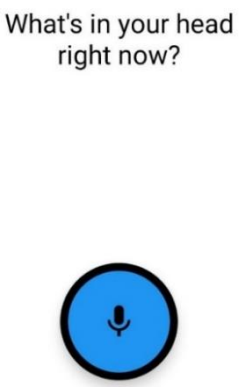
Slika 3.4 Prikaz dijagrama toka aplikacije

3.3.1. Korištenje aplikacije

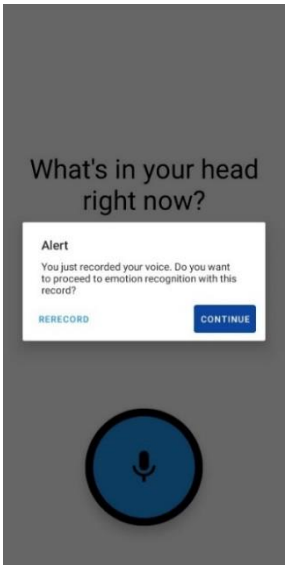
Pri pokretanju aplikacije pokazuje se početni zaslon (Sl. 3.5) koji prikazuje meni za navigaciju između modula funkcionalnosti za raspoznavanje emocija i modula za vizualizaciju već raspoznatih emocija.

Slika 3.5 Početni zaslon

Za početak procedure raspoznavanja korisnikove emocije, potrebno je snimiti glasovni odgovor na postavljeno pitanje držeći gumb za snimanje (Sl. 3.6). Završetak snimanja se obavlja otpuštanjem gumba za snimanje. Nakon snimljenog glasovnog zapisa, korisnik ima opciju (Sl. 3.7) ponovno snimiti glasovni zapis ili prijeći na korak raspoznavanja emocije.

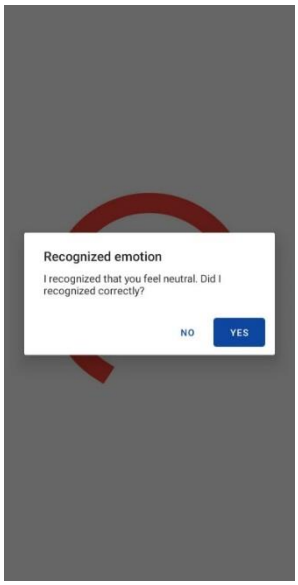


Slika 3.6 Fragment za snimanje glasovnog zapisa

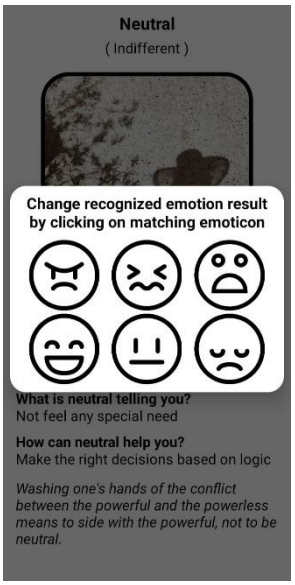


Slika 3.7 Prikaz dijaloga za odlučivanje ponovnog snimanja glasovnog zapisa

Nakon prihvaćenog snimljenog glasovnog odgovora i dočekanog odgovora s vanjskog servisa, prikazuje se dijalog gdje korisnik prihvaća raspoznatu emociju (Sl. 3.8). Ukoliko korisnik smatra da je sustav pogriješio, ima mogućnost odabrati neku drugu emociju, također putem dijalog prikaza (Sl. 3.9).

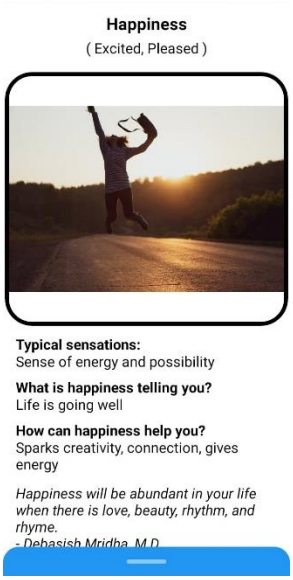


Slika 3.8 Prikaz dijaloga s raspoznatom emocijom



Slika 3.9 Prikaz dijaloga s odabirom emocije

Nakon validacije emocije, prikazuje se zaslon s informacija o toj emociji (Sl. 3.10). Prikazuje se fotografija, citat i tekst u kontekstu emocije. Osim navedenog zaslona, na dnu ekrana prikazana je ladica, koju je moguće prikazati povlačenjem od dolje prema gore. Otvorena ladica prikazuje informacije o *Solfeggio* frekvencijama i omogućuje reprodukciju jedne od 9 frekvencija koja je povezana konkretno uz raspoznatu emociju (Sl. 3.11). Isti gumb se koristi za reprodukciju i zaustavljanje audiozapisa, te se naravno gumb ažurira ovisno o stanju puštanja audiozapisa. Puštanjem audiozapisa (koji može biti reproduciran u pozadini iako je aplikacija zatvorena), Android sustav kreira notifikaciju s funkcijom zaustavljanja audiozapisa.



Slika 3.10 Prikaz zaslona s odgovarajućim informacijama o raspoznatoj emociji

Slika 3.11 Prikaz zaslona (otvorene ladice) o *Solfeggio* frekvencijama

Klikom na gumb (na početnom zaslonu) za prikaz pohranjenih raspoznatih emocija, prikazuje se vizualizacija emocija po danu (Sl. 3.12 i Sl. 3.13). Zaslon sadrži horizontalnu listu od 7 dana kojom se moguće pomicati povlačenjem (engl. *swipe*) s lijeva na desno. Svaki dan ima vertikalnu listu kojom se moguće pomicati povlačenjem od dolje prema gore.

Slika 3.12 Prikaz zaslona s vizualizacijom raspoznatih emocija

Slika 3.13 Prikaz dana u kojemu nema raspoznatih emocija

3.3.2. Ograničenja i moguća poboljšanja programskog rješenja

Eksperimentalnom analizom su se utvrdile performanse pojedinih modela te je SVM klasifikator učen nad EmoDB bazom podataka odabran za aplikacijsku primjenu. Iako, za razliku od ostalih modela je postigao bolje rezultate, performanse odabranog modela nisu savršene za korištenje u aplikaciji te se zbog toga implementirala validacija emocija u kojoj korisnik ima opciju odabrati ispravnu emociju u slučaju netočnog raspoznavanja emocije. Prema [20, 36], moguća poboljšanja performansi klasifikacijskog modela su korištenje baze podataka s emocijama prirodnog govora umjesto baze s glumljenim emocijama. Kako, prema [42], klasifikatori dubokog učenja zahtijevaju veću količinu podataka, na performanse klasifikatora dubokog učenja utjecalo bi povećanje količine podataka te povećanje raznolikosti podataka. Također, vezano uz poboljšanje performansi klasifikatora dubokog učenja, drugačije arhitekture konvolucijske neuronske mreže te korištenje tehnike prijenosa znanja (engl. *transfer learning*) predstavljaju budući smjer istraživanja ovog rada. Osim klasifikacijskog modela, moguće je i poboljšati Android aplikaciju dodatnim sadržajem, većom razinom interaktivnosti te implementacijom funkcionalnosti kojom bi aplikacija prepoznala loše emocionalno stanje korisnika i time reagirala sadržajem koje bi oraspoložilo korisnika.

4. ZAKLJUČAK

Raspoznavanje emocija predstavlja kompleksan proces prepoznavanja emocionalnog stanja osobe s kojom se stupilo u komunikaciju te je danas neizbježna vještina u cilju poboljšanja međuljudskih odnosa. Razvojem pametnih uređaja i sustava s kojima čovjek danas stupa u interakciju, računalno raspoznavanje emocija igra značajnu ulogu kako bi se unaprijedila interakcija između stroja i osobe u svrhu kvalitetnijeg korisničkog iskustva. Danas predstavlja popularan problem koji se još usavršava i nastoji se prijeći preko prepreka poput različitosti kultura, rasa i osobnosti.

U ovome radu ostvarena je usporedba između klasifikatora klasičnog strojnog učenja i klasifikatora dubokog strojnog učenja te su njihove performanse uspoređene na dvije baze. Performanse klasifikatora su ocijenjene mjerama kvalitete kao što su točnost, preciznost, odziv i F1-mjera. Klasifikatori su uspoređivani srednjom vrijednošću i standardnom devijacijom navedenih mjera kvaliteta na makro razini i uspoređivani su po emocijama F1-mjerom. Analizom eksperimenta utvrđeno je kako je najbolje i najstabilnije rezultate postigao SVM klasifikator nad EmoDB bazom te je ovaj model odabran za aplikacijsku primjenu. Također, analizom je utvrđeno kako klasifikatori trenirani nad EmoDB bazom postižu bolje rezultate i kako F1-mjere variraju za pojedine emocije. Istrenirani model je izložen kao web usluga pomoću Python okvira Flask i iskorišten u Android aplikaciji Know yourself pomoću Retrofit biblioteke za mrežnu komunikaciju. Razlog postavljanja modela kao web usluge umjesto ugrađivanja u aplikaciju leži u praktičnosti ažuriranja modela u budućnosti.

Rezultati modela korištenog u aplikaciji nisu savršeni te se zbog toga u aplikaciji korisniku nudi validacija raspoznate emocije u svrhu boljeg korisničkog iskustva. Prema [20, 36], model za praktičnu primjenu bi se mogao poboljšati korištenjem baze podataka s emocijama prirodnog govora umjesto baze s glumljenim emocijama. Također, povećanje količine i raznolikosti podataka bi utjecalo na performanse modela. Što se tiče aplikacije, moguća unaprjeđenja su implementacija dodatnog sadržaja u vezi raspoznate emocije koja bi omogućila višu razinu interaktivnosti i implementacija reakcije aplikacije ukoliko je korisnik u lošem emocionalnom stanju neko vrijeme.

LITERATURA

[1] M. Lamza – Maronić i J. Glavaš, Poslovno komuniciranje, Hrvatska, 2008.

[2] B. Farnsworth, How to Measure Emotions and Feelings (And the Difference Between Them), imotions.com, 2020., dostupno na https://imotions.com/blog/difference-feelings-emotions/ , [28. travnja 2022.]

[3] J. Freedman, Emotions, Feelings and Moods: What's the Difference?, 6seconds.org, dostupno na https://www.6seconds.org/2017/05/15/emotion-feeling-mood/ , [28. travnja 2022.]

[4] Y. Williams., Robert Plutchik's Wheel of Emotions, study.com, 2015., dostupno na https://study.com/academy/lesson/robert-plutchiks-wheel-of-emotions-lesson-quiz.html , [28. travnja 2022.]

[5] Z. Jamaludin, An Algorithm to Define Emotions Based on Facial Gestures as Automated Input in Survey Instrument, Malaysia, Advanced Science Letters , br. 10, sv. 22, str. 2889-2893, listopad, 2016.

[6] M. Pogosyan i J. B. Engelman, How We Read Emotions from Faces, Frontiers for Young Minds, br. 11, sv. 5, travanj, 2017.

[7] Health Jade, Auditory cortex, Health Jade Team, dostupno na https://healthjade.net/auditory-cortex/ , [28. travnja 2022.]

[8] S. Aman, Recognizing Emotions in Text, Ottawa-Carleton Institute for Computer Science School of Information Technology and Engineering University of Ottawa, 2007., dostupno na http://saimacs.github.io/pubs/2007-MS-Thesis.pdf [28. travnja 2022.]

[9] P. J. Lavrakas, Encyclopedia of Survey Research Methods, SAGE Publications, SAD, 2008.

[10] J. Sliwa, Best Way to Recognize Emotions in Others: Listen, American Psychological Association, 2017., dostupno na https://www.apa.org/news/press/releases/2017/10/emotions-listen [28. travnja 2022.]

[11] B. N. Reyes, S. C. Segal i M. C. Moulson, An investigation of the effect of race-based social categorization on adults' recognition of emotion, PLOS One, br. 2, sv. 13, veljača, 2018.

[12] F. A. Acheampong, C. Wenyu i H. Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities, Engineering Reports, br. 7, sv. 2, svibanj, 2020.

[13] C.Vinola i K.Vimaladevi, A Survey on Human Emotion Recognition Approaches, Databases and Applications, Electronic Letters on Computer Vision and Image Analysis, br. 14, sv. 2, str. 24-44, prosinac, 2015.

[14] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch i M.R. Wróbel, Emotion recognition and its applications, Advances in Intelligent Systems and Computing, Springer International Publishing Poland, sv. 300, str. 51-62, lipanj, 2014.

[15] M. Chan, This AI reads children's emotions as they learn, dostupno na https://edition.cnn.com/2021/02/16/tech/emotion-recognition-ai-education-spc-intl-hnk/index.html , [28. travnja 2022.]

[16] K. Vemou i A. Horvath, Facial Emotion Recognition, European Data Protection Supervisor, 2021., dostupno na https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-12021-facial-emotion-recognition_en , [26. kolovoza 2022.]

[17] N. Mehendale, Facial emotion recognition using convolutional neural networks (FERC), SN Applied Sciences, sv. 2, str. 2523-3971, veljača, 2020.

[18] D. Nautiyal, ML | Underfitting and Overfitting, GeeksforGeeks, 2022., dostupno na https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning , [11.5.2022.]

[19] V. V. Kamble , R. R. Deshmukh , A. R. Karwankar , V. R. Ratnaparkhe i S. A. Annadate, Emotion Recognition for Instantaneous Marathi Spoken Words, Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, sv. 2, str. 335-346, Indija, 2014.

[20] M. B. Akçaya i K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, sv. 116, str. 56-76, siječanj, 2020.

[21] Multidisciplinary Media & Mediated Communication, Acted Emotional Speech Dynamic Database – AESDD, Multidisciplinary Media and Mediated Communication, dostupno na http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition/ , [28. travnja 2022.]

[22] D. Cooper, CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), The Open Knowledge Foundation, dostupno na https://github.com/CheyneyComputerScience/CREMA-D , [28. travnja 2022.]

[23] G. Costantini, I. Iaderola, A. Paoloni i M. Todisco, EMOVO Corpus: an Italian Emotional Speech Database, European Language Resources Association (ELRA), 2014., dostupno na http://voice.fub.it/activities/corpora/emovo/index.html , [28. travnja 2022.]

[24] F. Burkhardt, M. Kienast, A. Paeschke i B. Weiss., Berlin Database of Emotional Speech, Technical University of Berlin, 1999., dostupno na http://emodb.bilderbar.info/index-1280.html , [28. travnja 2022.]

[25] S. R. Livingstone i F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), zenodo.org, 2018., dostupno na https://zenodo.org/record/1188976#.YhkR0OjMLRa , [28. travnja 2022.]

[26] University of Toronto, Toronto emotional speech set (TESS), University of Toronto, 2010., dostupno na https://tspace.library.utoronto.ca/handle/1807/24487 , [28. travnja 2022.]

[27] R. Lederman, Speech Emotion Recognition - Signal Preprocessing (1), Raphaël Lederman, 2019., dostupno na https://raphaellederman.github.io/articles/audioprocessing/#signal-preprocessing , [28. travnja 2022.]

[28] M. El Ayadi, M. S. Kamel i F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, br.3, sv. 44, str. 572-587, ožujak, 2011.

[29] H. Teager, Some observations on oral air flow during phonation, IEEE Transactions on Acoustics, Speech, and Signal Processing, IEEE, br. 5, sv. 28, str. 599-601, listopad, 1980.

[30] J. Kaiser, On a simple algorithm to calculate the ‘energy’ of the signal, International Conference on Acoustics, Speech, and Signal Processing, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), sv. 1, str. 381-384, SAD, 1990.

[31] L. Roberts, Understanding the Mel Spectrogram, Medium, 2020. dostupno na https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53 , [28. travnja 2022.]

[32] V. Velardo, AudioSignalProcessingForML, dostupno na https://github.com/musikalkemist/AudioSignalProcessingForML , [28. travnja 2022.]

[33] T. J. Khdour, A. A. Ahmad, S. K. Alqrainy i M. Alkoffash, Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques, Research Journal of Applied Sciences, Engineering and Technology, sv. 7, str. 5082-5097, lipanj, 2014.

[34] U. Shrawankar i Dr. V. Thakare, Techniques for feature extraction in speech recognition system : a comparative study, International Journal Of Computer Applications In Engineering, sv. 2, str. 412-418, svibanj, 2013.

[35] P. T. Krishnan, A. N. J. Raj i V. Rajangam, Emotion classification from speech signal based on empirical mode decomposition and non-linear features, Complex & Intelligent Systems, br. 2, sv. 7, str 1919–1934, veljača, 2021.

[36] B. J. Abbaschian, D. Sierra-Sosa i A. Elmaghraby, Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models, Sensors, br. 4, sv. 21, str. 1-27, SAD, veljača, 2021.

[37] S. Asiri, Machine Learning Classifiers, Towards Data Science, 2018., dostupno na https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 , [28. travnja 2022.]

[38] A. Chakure, Random Forest Classification, Medium, 2019., dostupno na https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0 , [28. travnja 2022.]

[39] Javatpoint, Support Vector Machine Algorithm, Javatpoint, dostupno na https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm , [28. travnja 2022.]

[40] Kathrin Melcher, A Friendly Introduction to [Deep] Neural Networks, Knime, 2021., dostupno na https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks , [28. travnja 2022.]

[41] R. M. S. de Oliveira, R. C. F. Araújo, F. J. B. Barros, A. Paranhos Segundo, R. F. Zampolo, W. Fonseca i V. Dmitriev, A System Based on Artificial Neural Networks for Automatic Classification of Hydro-generator Stator Windings Partial Discharges, Journal of Microwaves, Optoelectronics and Electromagnetic Applications, br. 3, sv. 16, str. 628-645, Brazil, studeni, 2017.

[42] E. Kavlakoglu, AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?, IBM, 2020., dostupno na https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks , [28. travnja 2022.]

[43] U. Sinha, Convolutions: Image convolution examples, AI Shack, 2017., dostupno na https://aishack.in/tutorials/image-convolution-examples/ , [26.5.2022.]

[44] A. Choulwar, The Art of Convolutional Neural Network, Medium, 2019., dostupno na https://medium.com/@achoulwar901/the-art-of-convolutional-neural-network-abda56dba55c , [26. travnja 2022.]

[45] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, SAD, 2019.

[46] M. Yani, S, Si. M. T. Budhi Irawan i C. Setianingsih, Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail, Journal of Physics: Conference Series, br. 1, sv. 1201, str. 012052, svibanj, 2019.

[47] S. Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, Towards Data Science, 2016., dostupno na https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 , [28. travnja 2022.]

[48] T. Murallie, 3 Ways to Deploy Machine Learning Models in Production, Towards Data Science, 2021., dostupno na https://towardsdatascience.com/3-ways-to-deploy-machine-learning-models-in-production-cdba15b00e , [28. travnja 2022.]

[49] P. Boersma, J. Ip i T. Gojani, Vokaturi - Android Library, Vokaturi, 2016., dostupno na https://github.com/alshell7/vokaturi-android , [28. travnja 2022.]

[50] Google, AffdexMe , Google, 2017., dostupno na https://play.google.com/store/apps/details?hl=en&id=com.affectiva.affdexme , [28. travnja 2022.]

[51] Google, Emotimeter - Emotion detector, Google, 2020., dostupno na https://play.google.com/store/apps/details?id=com.reaimagine.josem.emotimeter_facial emotionrecognizer , [28. travnja 2022.]

[52] Google, Group Emotion Recognition - Detect Face Expression, Google, 2019., dostupno na https://play.google.com/store/apps/details?id=com.hanuman.groupemotionrecognition , [28. travnja 2022.]

[53] C. T. Paddon, Solfeggio Tones Frequencies, Quantum Life Educational Newsletter, 2012., dostupno na http://shinewithlight.com/wp-content/uploads/2013/01/Solfeggio.pdf , [30. kolovoza 2022.]

[54] J. Robert, Pydub, PyPI.org, 2016., dostupno na https://github.com/jiaaro/pydub , [5.5.2022.]

SAŽETAK

U teorijskom dijelu rada opisan je problem raspoznavanja emocija. Dan je prikaz modela emocija te njihovo raspoznavanje i utjecaj na komunikaciju. Detaljnije je opisano računalno raspoznavanje emocija s korištenim postupcima, prikupljanje i predobrada podataka, izdvajanje značajki iz snimki govora i često korišteni klasifikatori. Opisana je mogućnost korištenja modela raspoznavanja emocija na Android platformi te su navedena već postojeća rješenja. U eksperimentalnom dijelu odrađen je postupak klasifikacije glasovnog zapisa. Na temelju spoznaja prikazanih u prvom dijelu rada, radi važnosti raspoznavanja emocija pri interakciji s računalnim sustavom, izrađeno je programsko rješenje za Android platformu. Programsko rješenje omogućuje raspoznavanje emocije iz snimke govore te prikaz informacija u svrhu unaprjeđivanja emocionalne inteligencije i prikaz raspoznatih emocija u posljednjih 7 dana. S tim ciljem, provedena je eksperimentalna analiza unutar koje su uspoređeni klasifikatori i baze podataka.

Ključne riječi: Android aplikacija, audio značajke, glasovno raspoznavanje emocija, klasifikacija, strojno učenje

ABSTRACT

Emotion recognition from recorded speech

In the theoretical part of the thesis, the problem of emotion recognition is described. The model of emotions together with their recognition and influence on communication are presented. Computer recognition of emotions with the procedures used, data collection and preprocessing, feature extraction from speech recordings and frequently used classifiers are described in detail. The possibility of using the emotion recognition model on the Android platform is described and already existing solutions are listed. In the experimental part, the voice recording classification procedure was performed. Based on the findings presented in the first part of the thesis, due to the importance of recognizing emotions when interacting with a computer system, a software solution for the Android platform was created. The software solution enables the recognition of emotions from speech recordings and also displays information about recognized emotion for the purpose of improving emotional intelligence. It also displays recognized emotions in the last 7 days. With this purpose, an experimental analysis was carried out in which classifiers and databases were compared.

Keywords: Android application, audio features, speech emotion recognition, classification, machine learning

ŽIVOTOPIS

Martin Zagoršćak rođen je 4. ožujka 1999. godine u Osijeku. Nakon završene Elektrotehničke i prometne škole Osijek upisuje Preddiplomski sveučilišni studij Računarstvo na Fakultetu elektrotehnike, računarstva i informacijskih tehnologija u Osijeku te potom upisuje diplomski studij Računarstvo smjer Programsko inženjerstvo. Stručnu praksu odradio je u tvrtki Factory na mjestu Android praktikanta. Bio je demonstrator na kolegiju Razvoj mobilnih aplikacija. Zainteresiran za razvoj Android aplikacija.

PRILOZI

1. „Raspoznavanje emocija iz zvučnih snimki govora“ u .docx formatu
2. „Raspoznavanje emocija iz zvučnih snimki govora“ u .pdf formatu
3. Izvorni kodovi:
   3.1. https://gitlab.com/zagi031/speech-emotion-recognition
   3.2. https://gitlab.com/zagi031/speech-emotion-recognition-api
   3.3. https://gitlab.com/zagi031/know-yourself