DIPLOMSKI RAD

Prepoznavanje emocija u govoru

Jovana Radakovic

RN 38/20

Beograd, decembar 2024.

Sadržaj

1. Uvod

Svet u kojem živimo oblikovan je ljudskom inteligencijom, koja je omogućila razvoj izuzetnih tehnoloških dostignuća. Sledeći veliki izazov trenutno je stvaranje veštačke inteligencije koja može da obavlja zadatke kao i ljudi. Razvoj veštačke inteligencije omogućava mašinama da percipiraju, uče, planiraju i rešavaju probleme. Kako bi tehnologija bila efikasnija, potrebno je da se mašine prilagode ljudskom okruženju, a ne obrnuto, kao što je to bio slučaj do sada. Za razvoj sofisticiranijih mašina potrebno je uključiti emocionalnu inteligenciju, jer emocije igraju ključnu ulogu u donošenju odluka i obavljanju zadataka.

Emocionalna inteligencija podrazumeva prepoznavanje i razumevanje emocionalnih stanja, što uključuje razvoj algoritama za prepoznavanje emocija, donošenje odluka u skladu sa njima i simulaciju emocionalnih reakcija. Prepoznavanje emocija, iako jednostavno za ljude, predstavlja izazov za algoritme veštačke inteligencije, jer emocije često nisu jasno definisane i mogu biti pomešane [1].

Jedan od osnovnih načina izražavanja emocija je govor, ali i izrazi lica, pokreti tela i fiziološki pokazatelji, kao što su puls ili aktivnost mišića [1].

S obzirom na brzi razvoj tehnologije, potreba za prirodnijom i intuitivnijom interakcijom između ljudi i računara postaje sve izraženija. Kako je govor najintuitivniji i najefikasniji oblik komunikacije za ljude, prirodno je očekivati da računari ne samo da razumeju govorne komande, već i da mogu interpretirati emocije koje se kriju iza tih komandi. U tom kontekstu, prepoznavanje emocija iz govora igra ključnu ulogu u unapređenju interakcije između ljudi i računara. Prepoznavanje emocija u govoru postalo je multidisciplinarno istraživačko polje koje obuhvata računarske nauke, psihologiju, fiziologiju, lingvistiku i umetnost.

Prepoznavanje emocija iz govora ima ključnu ulogu u različitim situacijama, uključujući krizne slučajeve, gde sistemi mogu detektovati strah ili paniku i prilagoditi svoje reakcije u skladu s tim. U svakodnevnom životu, emocionalno inteligentna mašina poboljšava korisničko iskustvo, omogućavajući personalizovane odgovore virtuelnih asistenata i četbotova. Takođe, u oblastima kao što su zdravstvo i edukacija, automatska analiza emocija iz govora može pomoći u prepoznavanju stanja poput depresije, anksioznosti ili stresa, pružajući korisnicima adekvatnu

podršku. Na taj način, prepoznavanje emocija ne samo da unapređuje tehnološke sisteme, već doprinosi i boljim međuljudskim interakcijama [2].

1.1. Ciljevi rada

Cilj ovog rada je istražiti efikasnost različitih dubokih arhitektura neuronskih mreža u prepoznavanju emocija iz govora. Konkretno, rad ima sledeće ciljeve:

1. Analizirati i razumeti osnovne karakteristike govora koje sadrže emocionalne informacije.
2. Uporediti performanse nekoliko arhitektura (CNN, LSTM, VGG16) na skupu podataka RAVDESS. Razviti i evaluirati modele.
3. Identifikovati ograničenja postojećih modela i predložiti unapređenja za buduća istraživanja.

1.2. Struktura rada

Rad je organizovan u nekoliko povezanih poglavlja kako bi se na sistematičan način predstavila tema:

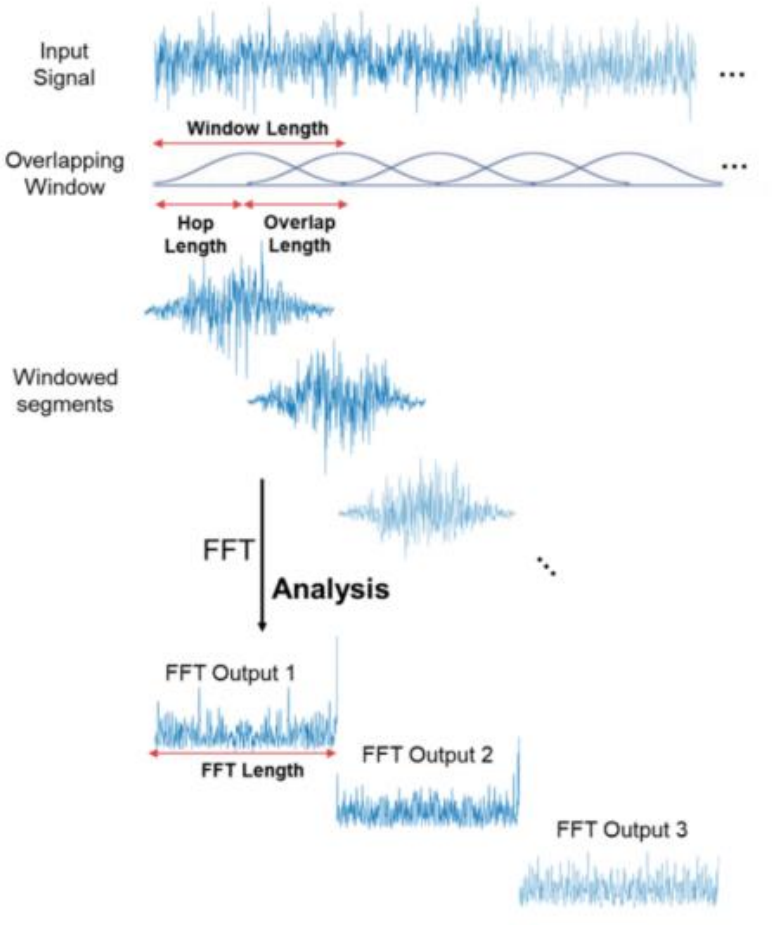
- Drugo poglavlje opisuje karakteristike govornog signala i metode za ekstrakciju karakteristika.
- Treće poglavlje objašnjava arhitekture izabranih modela.
- Četvrto poglavlje opisuje korišćeni skup podataka i tehnike pretprocesiranja koje su vršene nad tim podacima. Zatim preciznije karakteristike modela koji su objašnjeni u prethodnom poglavlju. Na kraju su prikazani rezultati eksperimenata i data je diskusija o njihovoj interpretaciji.
- Peto poglavlje rezimira rad, ističe ograničenja i predlaže pravce za buduće istraživanje.

2. Obeležja govornog signala

Neosporno je da su u snimljenom govornom signalu sadržane informacije o emocionalnom stanju govornika, pri čemu je neizostavni deo i lingvistička poruka, kao i informacije svojstvene za govornika, poput boje glasa, pola, starosti, ali i karakteristike prenosnog kanala. Klasifikacija emocionalnih stanja upotrebom neizmenjenog govornog signala ne bi bila efikasna, stoga je potrebno izvršiti izdvajanje obeležja (karakteristika) govornog signala koja što vernije oslikavaju emocionalna stanja uz potiskivanje ostalih sadržaja. Mnoštvo obeležja je izučavano, ali se nije došlo do univerzalnog rešenja, pri tome je akcenat ispitivanja prevashodno bio na akustičkim obeležjima, a u nekim studijama su kao dopuna korišćena i lingvistička obeležja [3]. Iako ne postoji univerzalna karakteristika kao rešenje svih klasifikacionih problema, na osnovu istraživanja dosadašnjih radova na tu temu može se reći da se trenutno najviše koriste mel-frekvencijski kepstralni koeficijenti (eng. Mel-Frequency Cepstral Coefficient - MFCC).

2.1 Mel-Spektrogram

Audio signal se može razložiti na sinusne i kosinusne talase koji formiraju originalni signal. Frekvencije i amplitude ovih reprezentativnih talasa mogu se koristiti za konverziju ulaznog signala iz vremenskog u frekvencijski domen. Brza Furijeova transformacija (eng. Fast Fourier transform-FFT) je algoritam koji se može koristiti za izvođenje ove konverzije. Međutim, FFT se izvodi na jednom vremenskom prozoru ulaznog signala. Ako se frekvencije reprezentativnih signala menjaju tokom vremena (neperiodične), ta promena ne može biti uhvaćena kroz konverziju jednog vremenskog prozora. Korišćenje FFT-a preko više preklapajućih prozora može se koristiti za konstrukciju spektrograma koji predstavlja amplitudu reprezentativnih frekvencija dok se one menjaju tokom vremena [4]. Ovi koraci su slikovito prikazani na slici 1.



SLIKA 1: PROCES PREKLAPANJA, PROZOROVANJA I PRIMENE FFT NA ULAZNI SIGNAL

2.2 Mel-frekvencijski kepstralni koeficijenti

Mel-frekvencijski kepstralni koeficijenti (MFCC) su najčešće korišćena obeležja u obradi govornog signala i prepoznavanju govornika i emotivnih stanja. Osnovna ideja MFCC-a je da oponaša način na koji ljudski slušni aparat percipira različite frekvencije. Koraci izvlačenja MFCC iz govornog signala šematski su prikazani na slici 4.

Proces izračunavanja MFCC-a je sledeći:

- Podela signala na segmente

  Govorni signal se deli na kratke segmente (frejmove) da bi se analizirale promene u vremenu. Segmenti se uzimaju sa pomerajem od 10 do 15 ms.

- Primena Hammingovog prozora

  Na svaki segment primenjuje se Hammingov prozor kako bi se smanjili efekti diskontinuiteta na krajevima segmenata. Jednačina za Hamming prozor je sledeća:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), \text{n} = 0, \ldots, \text{N-1},$$

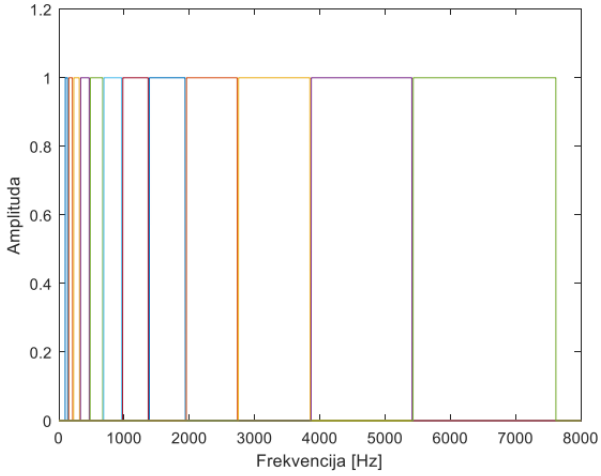
  gde je N broj uzoraka u segmentu.

- Transformacija u frekvencijski domen

  Koristi se diskretna Furijeova transformacija (DFT) za prelazak iz vremenskog u frekvencijski domen.

- Izračunavanje spektra snage

  Kvadriranjem magnituda frekvencijskog spektra dobija se spektar snage. Spektar se zatim deli na M opsega korišćenjem pravougaonih filtera prikazanih na slici 2.



SLIKA 2: BANKA PRAVOUGANIH FILTERA

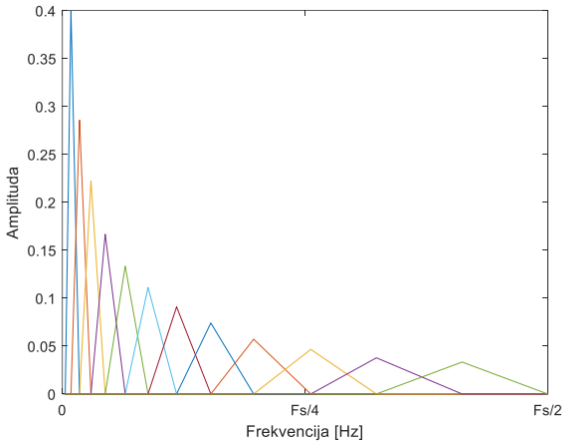
- Primena mel filter banke

  Formira se banka od M (M=12) trougaonih filtera (prikazani na slici 3) koji su ravnomerno raspoređeni na mel skali. Ovi filteri se primenjuju na spektar snage da bi se dobila raspodela energije na mel skali. Svaki filter definiše se formulom:

$$H_m(\phi) = \begin{cases} \dfrac{\phi - \phi_{b_{m-1}}}{\phi_{b_m} - \phi_{b_{m-1}}}, & \phi_{b_{m-1}} \le \phi \le \phi_{b_m} \\ \dfrac{\phi_{b_{m+1}} - \phi}{\phi_{b_{m+1}} - \phi_{b_m}}, & \phi_{b_m} \le \phi \le \phi_{b_{m+1}} \\ 0 & , \phi < \phi_{b_{m-1}} \ \text{V} \ \phi > \phi_{b_{m+1}} \end{cases}$$

gde je m = 1, …, M indeks filtra, a ϕ predstavlja diskretnu frekvenciju na mel-skali. Granične frekvencije $\phi_{b_0}, \ldots, \phi_{b_{m+1}}$ dele mel-skalu na M + 1 jednakih frekvencijskih opsega, pri čemu maksimalna frekvencija na mel-skali odgovara vrednosti $F_s/2$ na linearnoj (Hz) skali, gde je F_s učestalost odabiranja. Filtri se dalje transformišu u linearnu skalu zahvaljujući relaciji mel i linearne skale koja je opisana datom formulom:

$$\phi = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right), f \in \{\kappa Fs/N, \kappa = 0, \ldots, N/2\}$$



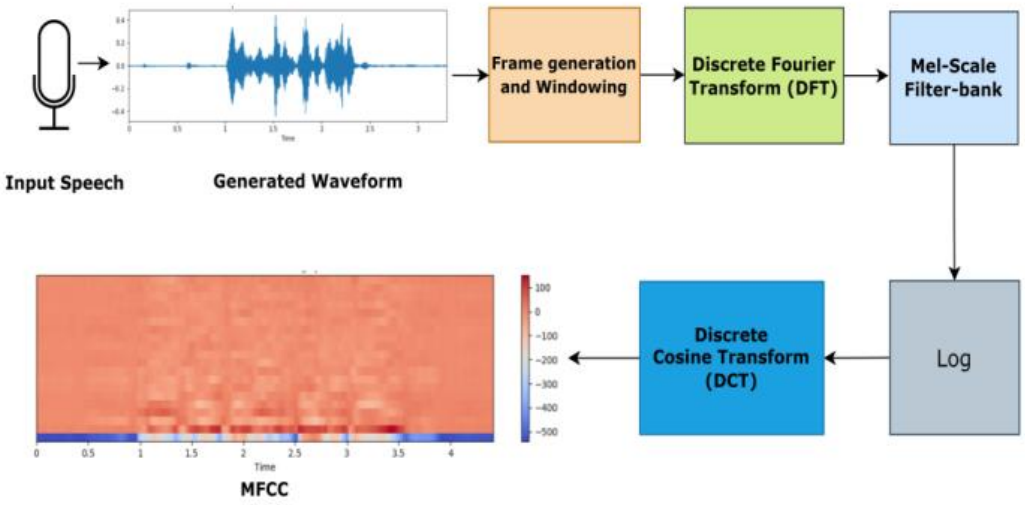
SLIKA 3: BANKA TROUGANIH FILTERA

- Logaritmovanje energije

  Logaritam energije za svaki filter koristi se kako bi se smanjio dinamički opseg i uvele perceptivne karakteristike.

- Diskretna kosinusna transformacija (DCT)

  Primenom DCT na logaritam energije dobijaju se MFCC koeficijenti. Ova transformacija omogućava kompaktnu reprezentaciju, gde su prvi koeficijenti najvažniji. Rezultat je M kepstralnih koeficijenata, gde se za vrednost uobičajeno bira M = 13 [5].



SLIKA 4: KORACI IZVLAČENJA MFCC IZ GOVORNOG SIGNALA

2.3 Koren srednje kvadratne amplitude u okviru (eng. Root Mean Square-RMS)

RMS predstavlja meru energije ili snage signala, i koristi se za kvantifikaciju amplitude zvučnog signala. RMS se računa kao kvadratni koren prosečne kvadratne vrednosti amplituda signala. RMS vrednost opisuje efektivnu vrednost signala, tj. njegovu prosečnu snagu tokom određenog perioda, i često se koristi za analizu jačine zvuka u vremenskom domenu.

Formula za računanje RMS je sledeća:

$$RMS_{\mathrm{t}} = \sqrt{\frac{1}{K} \cdot \sum_{\kappa = t \cdot K}^{(t+1) \cdot K - 1} s(\kappa)^2}$$

gde je t indeks trenutnog frejma (segmenta) signala, K je broj odbiraka (semplova) u jednom frejmu, κ je indeks odbirka u signalu, $s(\kappa)$ je amplituda signala za dato κ.

RMS se računa tako što se signal podeli na frejmove, svaki sa K odbiraka. Za svaki odbirak $s(\kappa)$ unutar frejma vrednost se kvadrira $s(\kappa)^2$. Zatim se računa zbir svih kvadriranih vrednosti. Taj zbir se deli sa ukupnim brojem odbiraka K i ovaj rezultat se korenuje kako bi se dobila efektivna vrednost amplituda u tom frejmu [6].

2.4 Stopa prolaska kroz nulu (eng. Zero-crossing rate - ZCR)

ZCR daje informaciju o broju promena znaka govornog signala za vreme trajanja jednog frejma, odnosno koliko puta signal pređe preko horizontalne ose. ZCR predstavlja grubu procenu učestalosti zastupljenih frekvencija u govornom signalu, tako da se može zaključiti da li je u pitanju zvučni govor, bezvučni govor ili pauza. Očekuju se relativno visoke vrednosti ZCR za bezvučni govor u odnosu na zvučni, dok u toku pauze vrednosti treba da budu bliske nuli (iako ambijentalni šum može da prouzrokuje jako visoke vrednosti ZCR) [6].

Za datu sekvencu odbiraka govornog signala $x(0), \ldots, x(N-1)$, gde je N broj uzoraka u segmentu a $x(\mathrm{n})$ je vrednost signala u trenutku n, ZCR je određen formulom:

$$ZCR = \frac{1}{2N} \sum_{n=1}^{N-1} |sign(x(n)) - sign(x(n-1))|,$$

pri čemu je

$$sign(x(n)) = \begin{cases} 1, & x(n) > 0 \\ -1, & x(n) \leq 0 \end{cases}.$$

3. Metodologija

Postupak prepoznavanja emocionalnog stanja govornika sadržanog u datoj govornoj sekvenci se sastoji iz dva koraka. Prvi korak podrazumeva izdvajanje obeležja govornog signala prema nekoj od metoda iz prethodnog poglavlja, dok je u drugom koraku potrebno doneti odluku o sadržanom emocionalnom stanju na osnovu izdvojenih obeležja. Različite tehnike su korišćene u klasifikaciji emocionalnih stanja, a najčešće su to: CNN (konvolucione neuronske mreže eng. Convolutional Neural Networks), LSTM (duga kratkoročna memorija eng. Long short-term memory), HMM (skriveni Markovljevi modeli, eng. Hidden Markov Models), GMM (model Gausovih mešavina, eng. Gaussian Mixture Models), SVM (metoda potpornih vektora, eng. Support Vector Machines) [7]. U nastavku će biti nešto više reči o klasifikatorskim metodama korišćenim u ovom radu, pri čemu su metode od izbora: CNN kao tradicionalan metod u obradi govora, LSTM koji se pokazao superiornim u mnogim zadacima prepoznavanja oblika i VGG16 model kao prethodno istreniran model primenjen na našim podacima.

3.1 Konvolucione neuronske mreže

Konvolucione neuronske mreže (CNN) su vrsta dubokih neuronskih mreža koje su posebno dizajnirane za obradu podataka sa grid strukturom, kao što su slike i video zapisi. One su postale standardni pristup u zadacima prepoznavanja slika i druge vizuelne obrade zbog svoje sposobnosti da automatski uče značajne karakteristike iz sirovih podataka.

Konvolucione neuronske mreže se sastoje od nekoliko ključnih slojeva koji su prikazani na slici 6 [8]:

1. **Konvolucioni sloj** (eng. Convolution layer):

   U konvolucijskom sloju vrši se linearna matematička operacija s matricama, zvana konvolucija, na način da se filterom (eng. Kernel) prolazi kroz ulazne podatke. Kada se pronađe određeni uzorak u podacima, on se preslikava u aktivacionu mapu. Na kraju konvolucijskog sloja nalazi se aktivaciona funkcija koja prilagođava raspon vrednosti

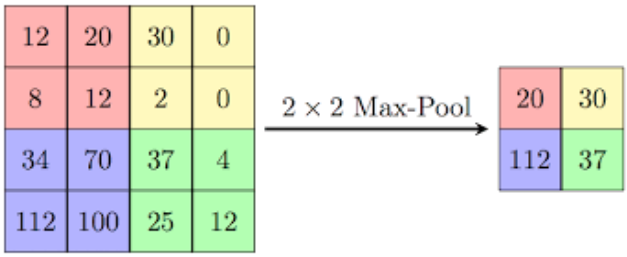
podataka. Aktivaciona funkcija koja se najčešće koristi je ReLU (eng. Rectified Linear Unit), defnisana kao:

$$ReLU(x) = \max(0,\ x)$$

gde je x ulaz u aktivacionu funkciju.

2. **Sloj sažimanja** (eng. Pooling layer):

   Ovaj sloj smanjuje dimenzionalnost aktivacionih mapa, čime se smanjuje broj parametara i računarska složenost mreže. Najčešće korišćene metode sažimanja su maksimalno sažimanje (max pooling), prikazano na slici 5 i prosečno sažimanje (average pooling).



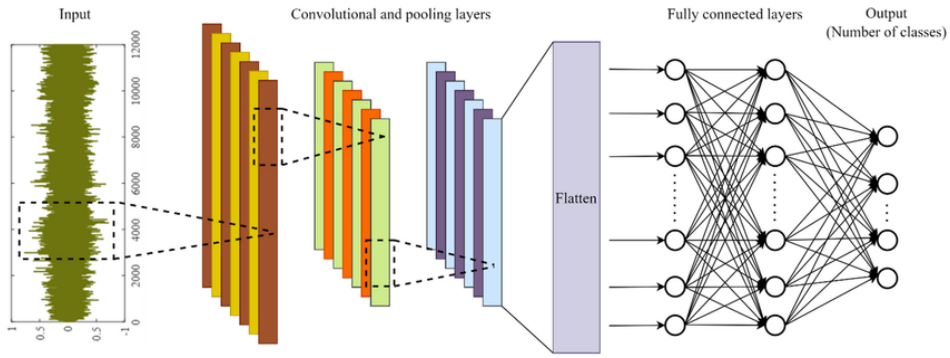
SLIKA 5: MAX-POOL OPERACIJA

3. **Potpuno povezani sloj** (eng. Fully connected layer):

   Na kraju mreže, ovaj sloj kombinuje informacije iz prethodnih slojeva i donosi konačnu odluku ili klasifikaciju na osnovu naučenih karakteristika.

Svi spomenuti slojevi mogu se kombinovati na različite načine kako bi se postigle što bolje performanse mreže. S obzirom na protok informacija kroz strukturu, konvolucijske neuronske mreže pripadaju tipu mreža sa propagacijom signala unapred. Kod neuronskih mreža sa propagacijom signala unapred (engl. Feedforward Neural Networks) ne postoje povratne petlje,

stoga informacije putuju samo unapred kroz mrežu, najpre kroz ulazne čvorove, zatim kroz skrivene čvorove i na kraju kroz izlazne čvorove [8].

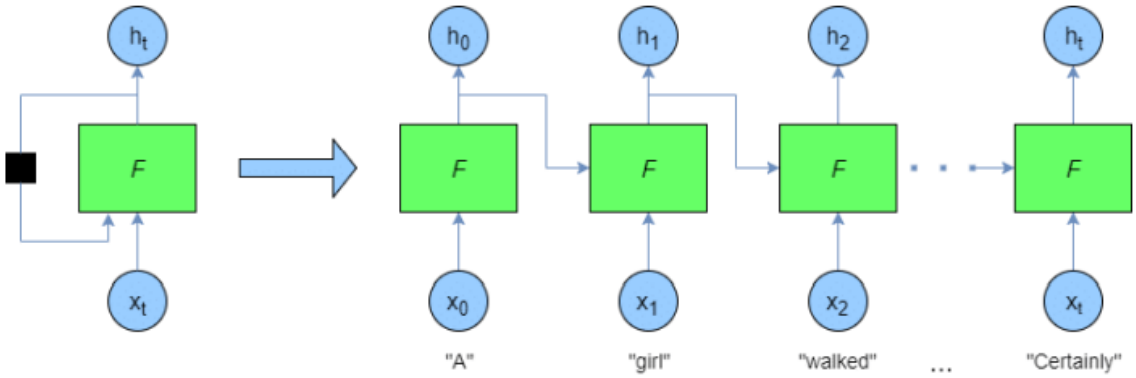


SLIKA 6: PRIKAZ 1D CNN

3.2 Duga kratkoročna memorija

3.2.1. Rekurentne neuronske mreže

Rekurentne neuronske mreže (engl. Recurrent Neural Networks - RNN) su posebne po tome što za razliku od mreža sa propagacijom signala unapred imaju povratnu vezu. To im omogućuje da izlaz iz prethodnog vremenskog koraka prosleđuju na ulaz u trenutni vremenski korak i na taj način utiču na rezultat.



SLIKA 7: ŠEMATSKA REPREZENTACIJA RAZVIJENOG RNN-A

Zbog toga je model povratne neuronske mreže dinamičan i nakon procesa učenja. Na slici 7 se može videti da mreža istovremeno za podatke uzima trenutni ulaz i izlaz iz prethodnog vremenskog koraka.

Ove mreže koriste se za učenje sekvencijalnih podataka kao što su tekst, zvuk i podaci sa senzora. Slika takođe prikazuje i „odmotanu“ mrežu kojom je demonstriran način unošenja sekvencijalnih podataka. Pamteći prethodne informacije dok obrađuju trenutne, ove mreže mogu povezati niz događaja. Problem je što nakon određenog broja sekvenci, mreža počne „zaboravljati“ starije podatke, što se naziva problem nestajućeg gradijenta (engl. vanishing gradient problem). Taj problem rešava naprednija verzija rekurentne mreže, a to je LSTM (engl. Long Short-Term Memory, duga kratkoročna memorija). Ona je sposobna pamtiti bitne informacije kroz mnogo koraka i zbog toga može dobro razumeti kontekst. Kao i kod konvolucijske neuronske mreže, LSTM mreža se sastoji od ćelija (Slika 8) koje čine osnovni gradivni blok modela [9].

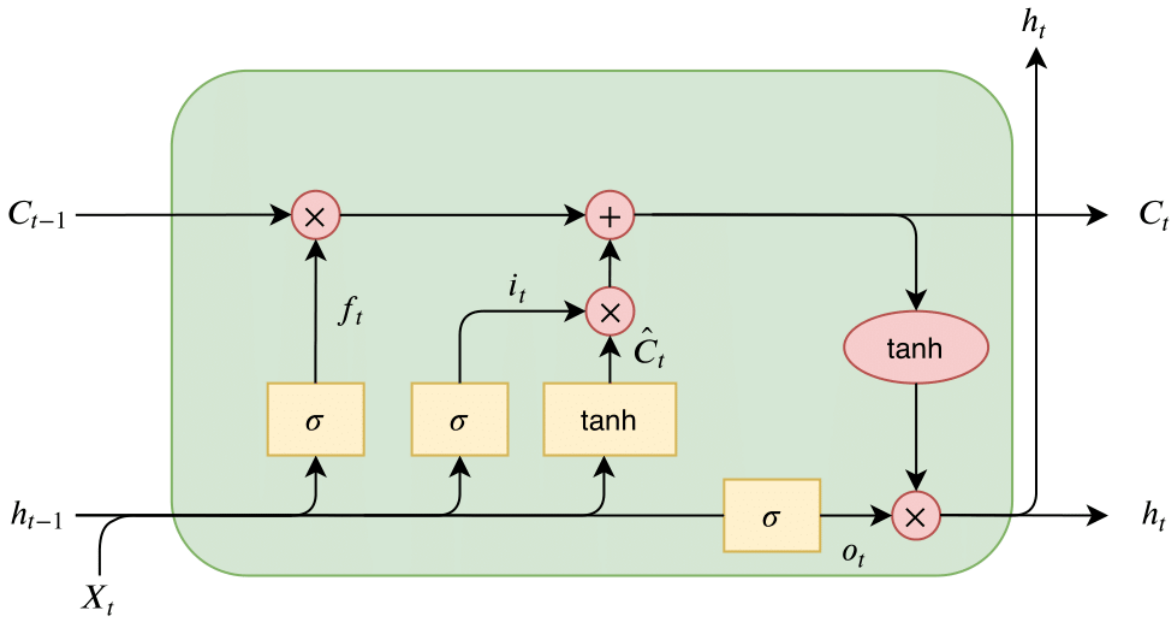
Svaka ćelija sadrži unutrašnje komponente koje omogućavaju kontrolu protoka informacija kroz četiri glavna dela:

- ulazna vrata (i_t)
  kontrolišu koje nove informacije se dodaju trenutnom stanju memorije,

- zaboravna vrata (f_t)
  odlučuju koje informacije iz prethodnog stanja treba ukloniti,
- stanje memorije (C_t)
  se ažurira kombinovanjem prethodne memorije i novih informacija,
- izlazna vrata (o_t)
  odlučuju šta će biti prosleđeno dalje kao trenutni izlaz mreže.

LSTM neuronska mreža sastoji se od više slojeva. Ulazni sloj prima ulazne sekvence. LSTM slojevi su slojevi sa više ćelija u kojima se obrađuju podaci. U LSTM sloju vrši se učenje odnosa među podacima kako bi se shvatio kontekst. Na kraju LSTM slojeva nalazi se gusti deo mreže (eng. dense (fully connected)) koji prevodi skrivena stanja u izlazne vrednosti.

Na kraju mreže nalazi se izlazni sloj, a u slučaju klasifikacije podataka, taj sloj sadrži Softmax aktivacijsku funkciju. Spomenuta funkcija dodeljuje verovatnoću pripadanja svakoj od mogućih klasa. Klasa s najvećom verovatnoćom uzima se kao konačna predikcija.



SLIKA 8: ARHITEKTURA JEDNE LSTM ĆELIJE

Ovo je arhitektura jedne LSTM ćelije [9]. Ulazi su X_t: trenutni ulazni podatak, h_{t-1}: skriveno stanje iz prethodnog vremenskog koraka i C_{t-1}: stanje ćelije iz prethodnog vremenskog koraka.

Zaboravna vrata f_t, prikazana kroz sigmoid (σ**)** i množenje ($\times$). Ova vrata odlučuju koje informacije iz C_{t-1} treba zaboraviti. Izračunava se formulom

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, X_t] + b_f\right)$$

Ulazna vrata i_t i kandidat za novo stanje $\hat{C}_t$, i_t je sigmoid funkcija koja određuje koje nove informacije treba dodati u stanje ćelije $\hat{C}_t$. Korišćenje tanh funkcije je za stvaranje kandidata za novo stanje. Ažurirano stanje se računa kao

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, X_t] + b_i\right)$$

$$\hat{C}_t = tanh\left(W_c \cdot [h_{t-1}, X_t] + b_c\right.$$

Ažurirano ćelijsko stanje C_t je kombinacija prethodnog stanja C_{t-1} i novih informacija:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t$$

Izlazna vrata o_t koriste sigmoid (σ) kako bi odlučila koje informacije iz trenutnog stanja ćelije C_t treba propuštati ka izlazu h_t. Izračunava se formulom

$$o_t = \sigma(W_0 \cdot [h_{t-1}, X_t] + b_0)$$

$$h_t = o_t \cdot \tanh\left(C_t\right)$$

Izlazi su h_t odnosno skriveno stanje (koristi se kao izlazni rezultat trenutnog koraka) i C_t odnosno ažurirano stanje ćelije koje se prenosi u sledećem vremenskom koraku.

3.3 VGG16

Model je razvijen od strane istraživača sa Univerziteta Oxford u okviru Visual Geometry Group (VGG), otuda i naziv VGG16 [9]. VGG16 je prvobitno razvijen za ImageNet Large Scale Visual Recognition Challenge 2014. i istrenirana na skupu podataka ImageNet. VGG16 je arhitektura koja se sastoji od 16 slojeva (od kojih je 13 konvolucijskih slojeva, 3 potpuno povezana sloja i jedan sloj za klasifikaciju). Glavni cilj ove arhitekture je da poveća dubinu modela kako bi poboljšala sposobnost mreže da uči složene obrazce. Glavni slojevi arhitektura čine sledeći i prikazani su na slici 9 [10] :

1. **Ulazni sloj**:

   Ulazni podaci su slike dimenzija 224×224×3, gde 224x224 predstavlja visinu i širinu slike, a 3 predstavlja tri kanala (RGB).

2. **Konvolucijski slojevi**:

   VGG16 koristi 3x3 konvolucijske filtere sa veličinom pomeraja **(**eng stride**)** 1 i veličinom dodate ivice (eng. padding) 1. To znači da se dimenzije slike ne menjaju nakon primene konvolucije. Svaki konvolucijski sloj koristi ReLU aktivacionu funkciju. Postoji ukupno 13 konvolucijskih slojeva u VGG16, raspoređenih u 5 blokova. Svaki blok počinje sa 2-3 konvolucijska sloja, a zatim sledi maksimalno sažimanje (eng. max-pooling) koje smanjuje dimenzije slike i omogućava modelu da se fokusira na osnovne karakteristike.

3. **Slojevi sažimanja**

   Nakon svakog niza konvolucijskih slojeva, koristi se metoda sažimanja, maksimalno sažimanje (eng. max-pooling) sa filterom 2x2 i veličinom pomeraja 2,

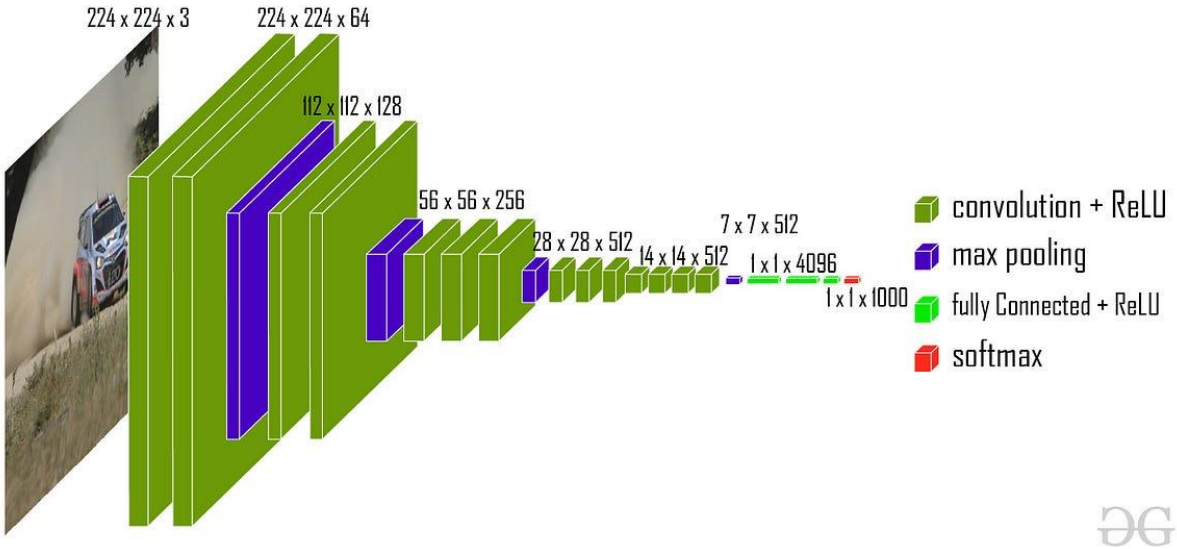
što smanjuje dimenzije ulaza za faktor 2. Ovo pomaže da se smanji broj parametara i složenost modela, dok se i dalje zadržavaju relevantne karakteristike.

4. **Potpuno povezani slojevi** :

   Nakon što su konvolucijski slojevi završeni, izlaz se ispravlja (eng. flatten) u jednostavan niz. Zatim, 3 potpuno povezana sloja obrađuju ove informacije. Prvi potpuno povezani sloj ima 4096 neurona, drugi takođe 4096 neurona, dok treći sadrži 1000 neurona koji odgovaraju broju klasa u ImageNet datasetu.

5. **Izlazni sloj**:

   Poslednji sloj koristi softmax funkciju aktivacije za klasifikaciju slika u 1000 klasa, što je specifično za ImageNet dataset. Za druge zadatke, broj izlaznih klasa može biti prilagođen.



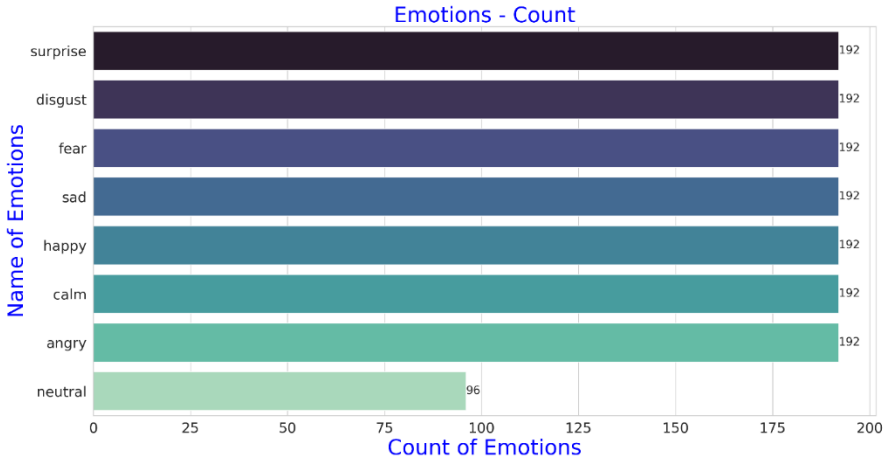
SLIKA 9: VGG16 ARHITEKTURA

4. Eksperimenti i Rezultati

4.1 Korišćeni podaci

RAVDESS (eng. Ryerson Audio-Visual Database of Emotional Speech and Song) je baza podataka koja je široko korišćena u istraživanjima prepoznavanja emocija u govoru i pesmama. Baza je kreirana 2018. godine od strane timova sa Ryerson University-a u Torontu, Kanada [11]. RAVDESS baza obuhvata osam osnovnih emocija, snimljeno od strane 24 glumca (12 žena i 12 muškaraca) koji izgovaraju dve rečenice "Kids are talking by the door" i "Dogs are sitting by the door". Emocije u bazi su sledeće i njihova distribucija u bazi podataka je prikazana na slici 10:

1. Neutralno (neutral)
2. Sreća (happy)
3. Tuga (sad)
4. Bes (angry)
5. Strah (fearful)
6. Gađenje (disgust)
7. Iznenađenje (surprised)
8. Mirnoća (calm)



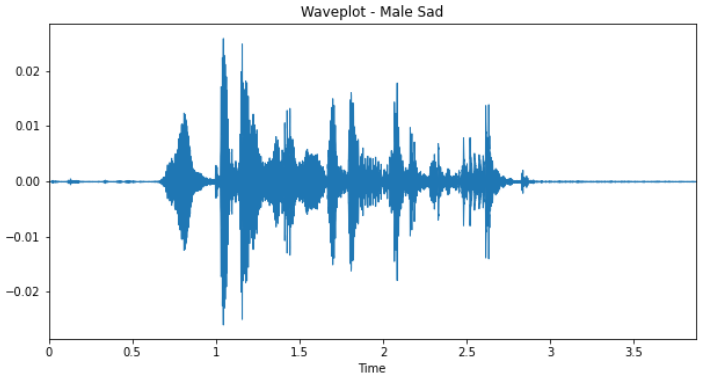
SLIKA 10: DISTRIBUCIJA RAZLIČITIH EMOCIJA U SKUPU PODATAKA

Podaci su podeljeni po glumcima s imenima datoteka “Actor_01”, “Actor_02” itd. Primer naziva jednog audio fajla iz datoteke je *02-01-06-01-02-01-12.mp4* a u nastavku je objašnjeno tačno značenje svakog obeležja.

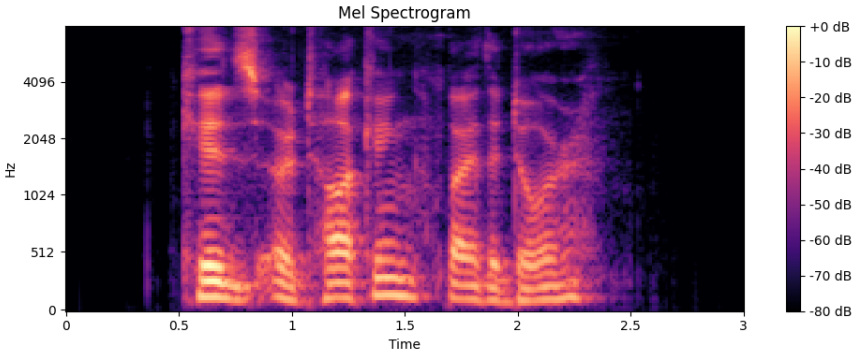
1. Modalitet (Modality): 01 za govor, 02 za pevanje,
2. Kanal (Channel): 01 za audio, 02 za audio-vizuelni zapis,
3. Emocija (Emotion): Kod za emociju (01-08),
4. Intenzitet (Intensity): 01 za normalni intenzitet, 02 za naglašeni intenzitet,
5. Izjava (Statement): 01 za prvu rečenicu, 02 za drugu rečenicu,
6. Ponavljanje (Repetition): 01 za prvu iteraciju, 02 za drugu iteraciju,
7. Glumac (Actor): ID glumca (01-24; muškarci imaju neparne brojeve, žene imaju parne brojeve).

4.2. Ekstrakcija karakteristika i tehnike augmentacije

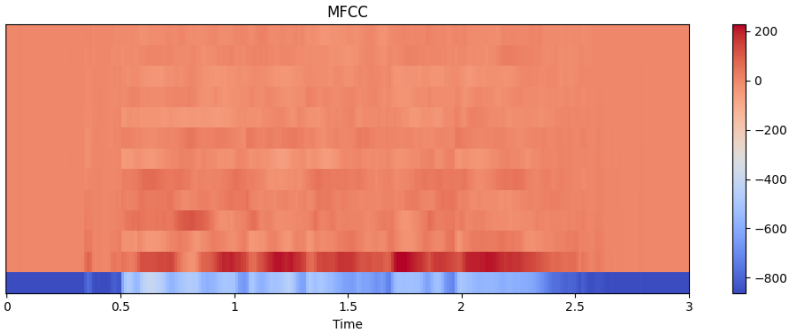
Iz svakog audio fajla potrebno je izvući karakteristike na osnovu kojih će dati modeli kasnije klasifikovati emocije. Prva karakeristika koja se koristi jeste mel spectrogram (Slika 12) druga karakteristika je MFCC (Slika 13). Obe su detaljnije objašnjene u drugom poglavlju. Na slici 11 je dat primer kako izgleda talasni oblik audio fajla bez ikakvih augmentacija, dok će kasnije biti prikazano kako se menja izgleda talasnog oblika kada se primene tehnike augemntacije (Slika 14, Slika 15, Slika 16).



SLIKA 11: TALASNI OBLIK AUDIO FAJLA EMOCIJE TUGE U VREMENSKOM DOMENU



SLIKA 12: MEL SPEKTROGRAM EMOCIJE TUGE

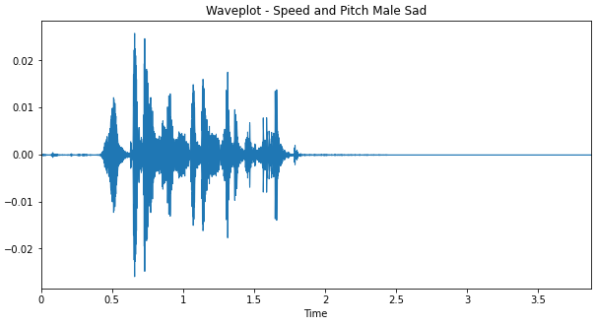


SLIKA 13: MFCC EMOCIJE TUGE

Na ove karakteristike se primenjuju tehnike augmentacije podataka kao što su:

1. Promena brzine i tona (eng. speed and pitch)

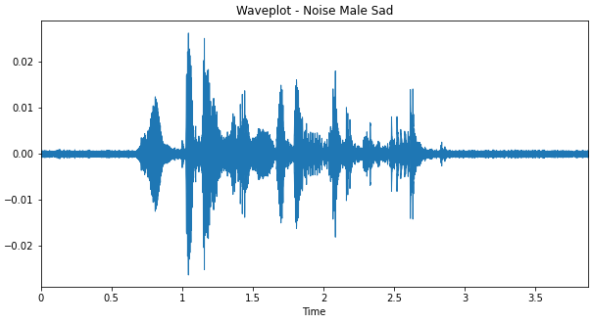
Ova tehnika podrazumeva ubrzavanje ili usporavanje zvučnog fajla. Kada se promeni brzina, menja se i ton zvuka. Zvuk može da se ubrza on tada ima viši ton ili da se uspori, tada ima niži ton. Ovo se često koristi da bi se simulirale različite brzine govora ili kako bi model bio otporniji na varijacije u govoru [12].



SLIKA 14: DODATO UBRZANJE I VIŠI TON NA TALASNI OBLIK EMOCIJE TUGE

4. Dodavanje šuma (eng. noise)

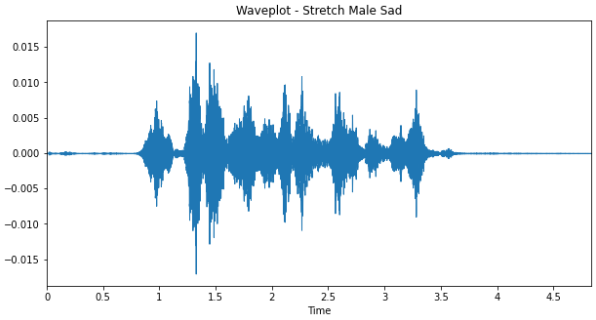
U ovu tehniku spada dodavanje pozadinskog šuma, poput šuma vetra, buke u prostoriji ili drugih neželjenih zvukova. Cilj je simulirati realne uslove u kojima se govor može snimati i povećati otpornost modela na šumove i učiniti ga robusnijim u stvarnim aplikacijama [12].



SLIKA 15: DODAT ŠUM NA TALASNI OBLIK EMOCIJE TUGE

5. Prerastezanje (eng. stretch)

Ova tehnika menja trajanje zvučnog fajla bez promene tona. Istezanje u vremenu podrazumeva produžavanje ili skraćivanje trajanja zvuka bez promene njegove visine. Ovo može pomoći da se model prilagodi različitim tempovima govora ili pevanja [12].



SLIKA 16: PRERASTEZANJE TALASNOG OBLIKA EMOCIJE TUGE

Ove tehnike se koriste kako bi se obogatio skup podataka i povećala raznolikost podataka, što na kraju poboljšava sposobnost modela da prepoznaje emocije. Tehnike augmentacije su primenjene samo na podatke za obučavanje.

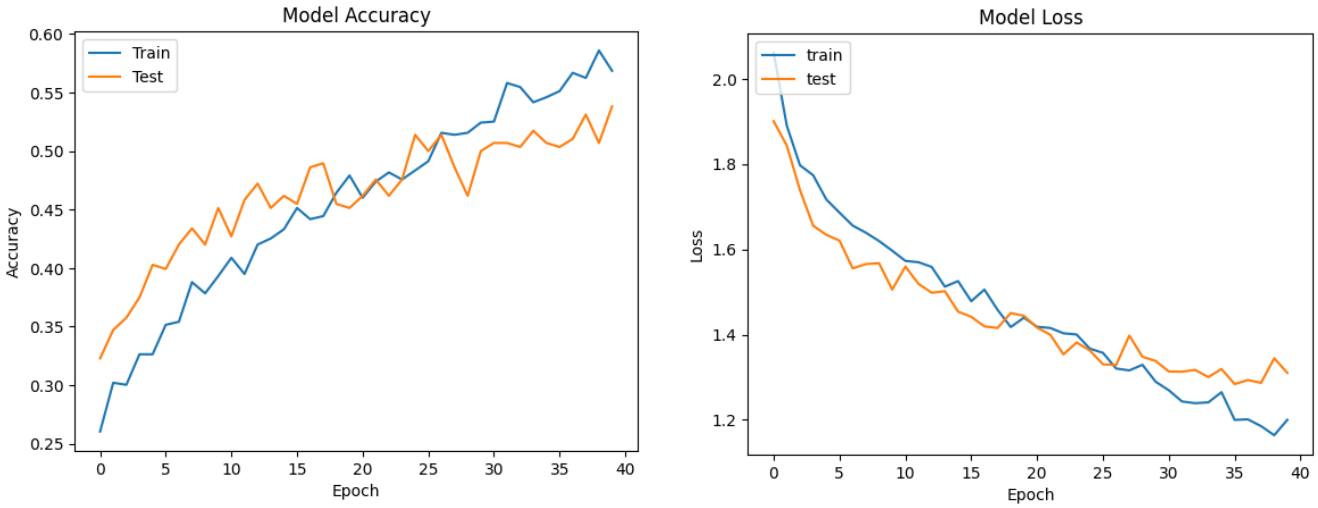
4.3 Arhitektura

4.3.1. Model konvolucijske neuronske mreže

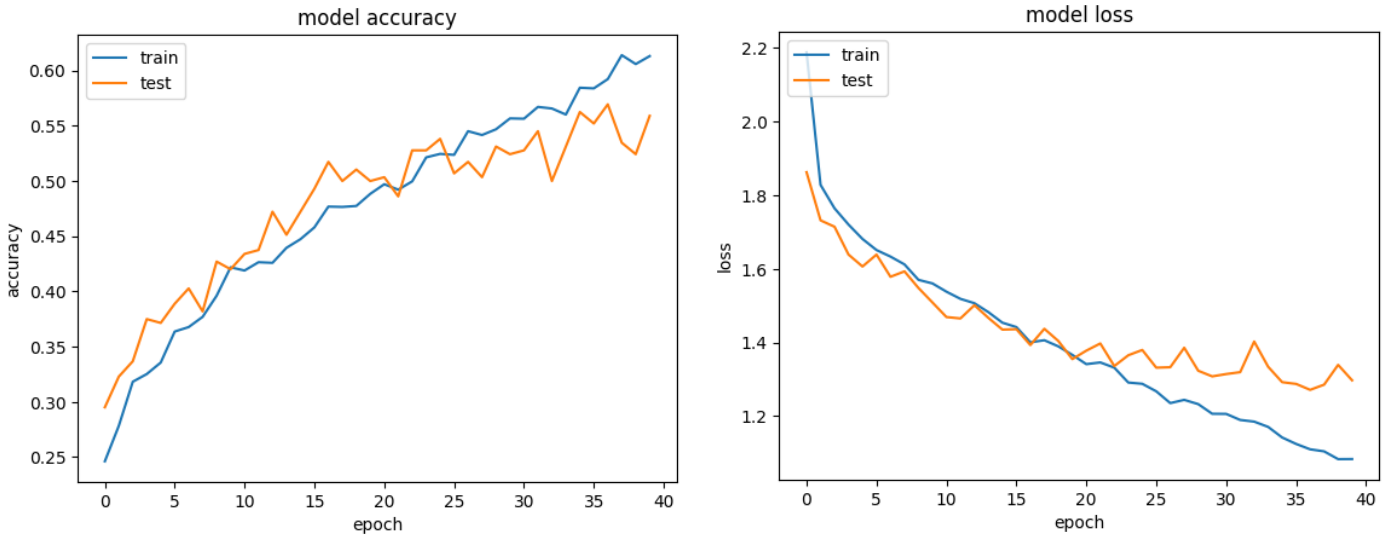
Prvi sloj primenjuje 1D konvuluciju na ulazne podatke koristeći 64 filtera. Veličina kernela je 10, što znači da se analiza vrši na segmentima od 10 uzoraka. Aktivacijska funcija je ReLU koja se koristi za uvođenje nelinearnosti. Drugi sloj je još jedna konvolucija sa većim brojem filtera – 128.

Regularizacija je L2 (sa parametrom 0.01) i koristi se za penalizovanje velikih težina i sprečavanje overfitinga dok je aktivacija svuda ReLu. Treći sloj je MaxPooling (veličina pool-a je 8), sloj koji vrši smanjenje dimenzionalnosti odabirom maksimuma iz svake grupe od 8 uzastopnih uzoraka. Cilj je zadržati najvažnije informacije i smanjiti količinu podataka za naredne slojeve. Naredni, četvrti sloj je Dropout (stopa je 0.4) koji se koristi za regularizaciju i smanjenje overfitinga tako što se 40% neurona nasumično isključi tokom treniranja. Peti sloj je konvolucija koja koristi 128 filtera, veličina kernela 10, ReLU aktivacija. Zatim još jedan MaxPooling (veličina pool-a 8). Nakon njega sledi Dropout (stopa je 0.4). Osmi sloj je Flatten, odnosno on pretvara višedimenzionalne podatke u jedan niz vrednosti (vektor) kako bi se podaci mogli predati gustim slojevima. Ovaj sloj je neophodan za prelazak sa konvolucionih na potpuno povezane slojeve. Deveti sloj je Dense (256 neurona, ReLU aktivacija). Ovaj sloj kombinuje informacije iz prethodnih slojeva i pronalazi složenije obrasce. Zatim sledi još jedan Dropout sloj (stopa je 0.4). Poslednji sloj, izlazni sloj je Dense sa 8 neurona (odgovara broju klasa emocija), aktivacijska funkcija sigmoid se koristi jer omogućava da se svaka klasa modelira kao nezavisna. Veličina batch-a je 32, broj epoha je 40. Dati su rezultati eksperimenata (grafovi preciznosti i troška za različite karakteristike govora - Slika 17, Slika 18, Slika 19 i Slika 20) koji su kasnije detaljnije prokomentarisani.

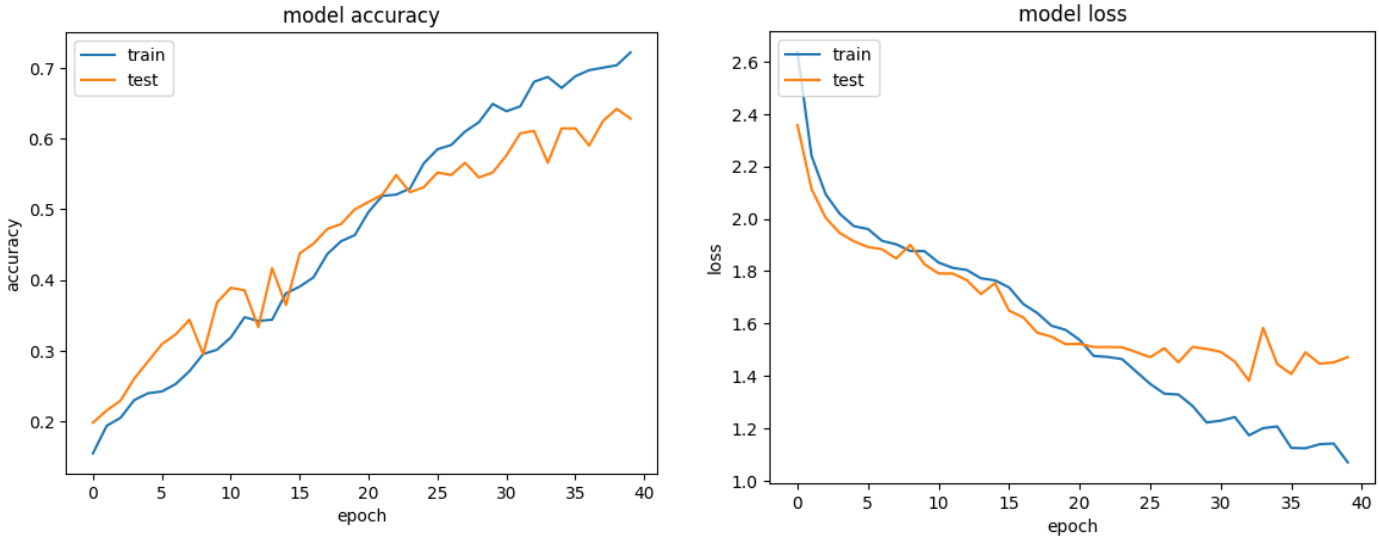
Jedina razlika u arhitekturi CNN za karakteristike mel spectrogram i MFCC jeste što kod CNN za MFCC, peti sloj je konvolucija koja koristi 128 filtera, veličine kernela 2.



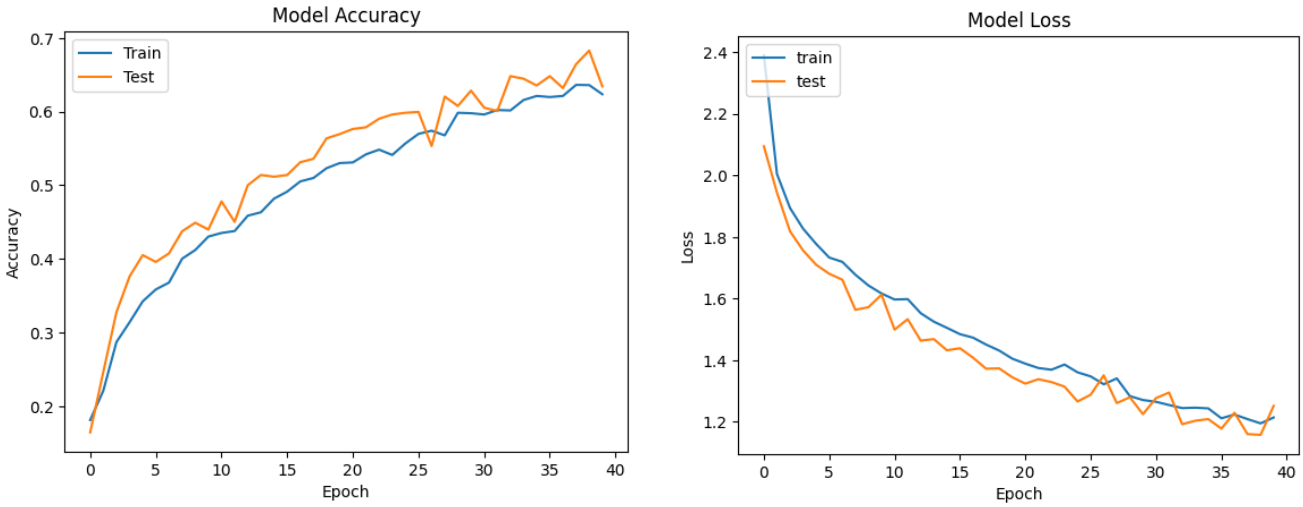
SLIKA 17: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) CNN ZA KARAKTERISTIKU LOG MEL SPECTROGRAM



SLIKA 18: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) CNN ZA KARAKTERISTIKU LOG MEL SPEKTROGRAM S AUGMENTACJIOM



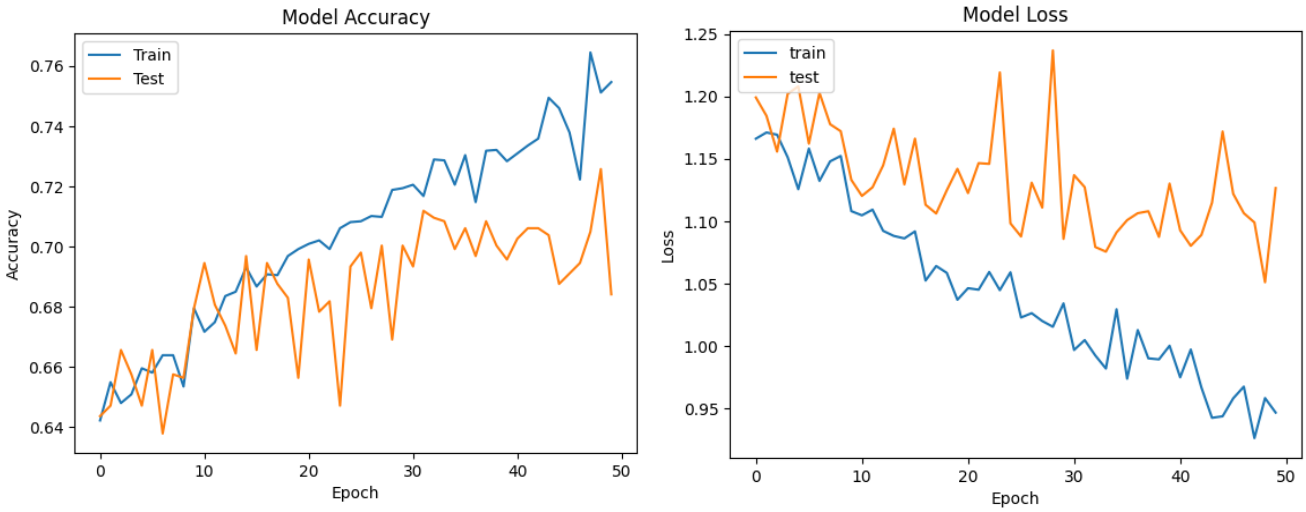
SLIKA 19: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) CNN ZA KARAKTERISTIKU MFCC



SLIKA 20: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) CNN ZA KARAKTERISTIKU MFCC S AUGMENTACIJOM

4.3.2 Model LSTM

Ova arhitektura se sastoji od pet slojeva, od kojih su tri LSTM sloja sa 256, 128 i 32 jedinice, redom. Svaki od ovih slojeva koristi dropout i rekurentni dropout vrednosti od 0.2 kako bi se smanjio efekat prekomernog obučavanja. U svakom LSTM sloju su dodati regularizatori jezgra (eng. kernel regularizers) kako bi se dodatno umanjila mogućnost prekomernog prilagođavanja. Nakon LSTM slojeva, model uključuje sloj za ravnanje (eng. flatten layer) koji konvertuje višedimenzionalni izlaz u jednodimenzionalni niz, što omogućava dalju obradu u gustim slojevima. Sledeći gusti sloj sadrži 256 neurona i koristi 'relu' aktivacionu funkciju, dok poslednji sloj predviđa izlaz među osam emocija koristeći 'softmax' aktivacionu funkciju. Rezultati eksperimenta, tačnije grafovi preciznosti i troška, prikazani su na slici 21 i kasnije su detaljnije prokomentarisani.



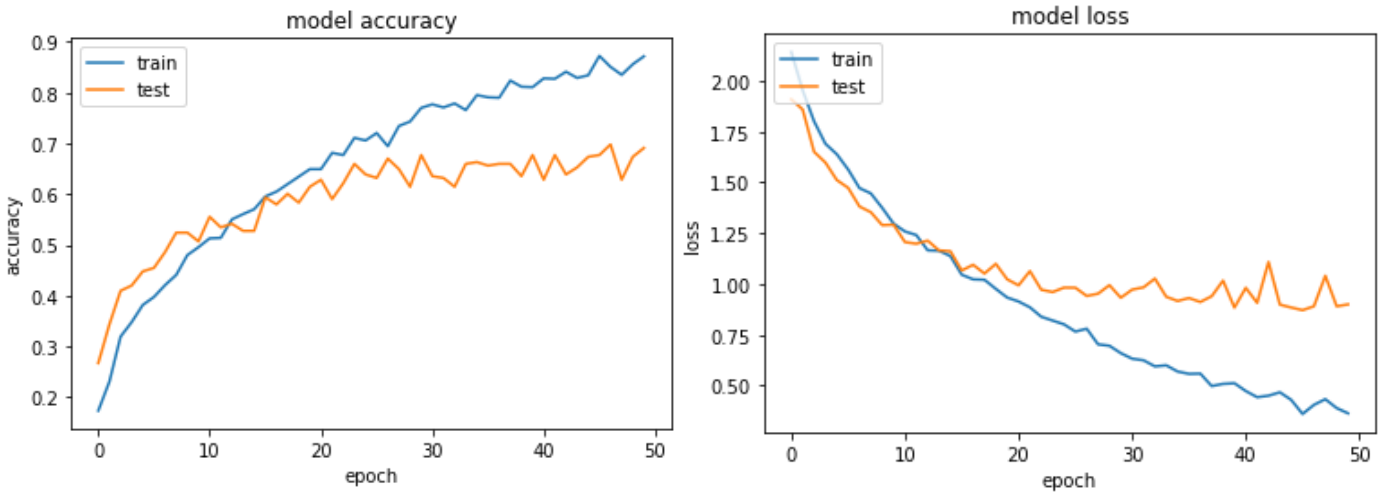
SLIKA 21: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) LSTM ZA KARAKTERISTIKU MFCC S AUGMENTACIJOM

4.4.3. Model VGG16

Učitavaju se unapred istrenirane težine sa ImageNet skupa podataka, dok se završni potpuno povezani sloj (eng. fully connected - FC) iz originalnog modela uklanja, čime se omogućava prilagođavanje specifičnom zadatku. Zamrzavaju se slojevi osnovnog modela da bi se sprečilo ažuriranje težina slojeva tokom treninga, jer su oni već obučeni za prepoznavanje opštih karakteristika (kao što su ivice, oblici itd.). Ulazne vrednosti su slike mel spectrograma govornih signala. Na vrh VGG16 modela dodati su prilagođeni slojevi kako bi se obavio zadatak klasifikacije u 8 klasa. Poravnanje (Flattening) pretvara izlaz poslednjeg konvolutivnog sloja u jednodimenzionalni vektor, pogodan za potpuno povezane slojeve [13].

Prvi je potpuno povezani sloj sa 512 neurona i ReLU aktivacionom funkcijom. Zatim sledi sloj koji "isključuje" nasumično odabrane neurone (0.5 tj. 50%) tokom treninga kako bi se smanjila zavisnost izlaza od ulaza. Sledi još jedan potpuno povezani sloj sa 256 neurona i ReLU aktivacijom. Završni sloj ima 8 neurona (po jedan za svaku klasu) i softmax aktivacionu funkciju,

koja generiše verovatnoće za svaku klasu. Na kraju se kombinuju VGG16 osnovni model i prilagođeni slojevi u jednu mrežu. Rezultati eksperimenta, tačnije grafovi preciznosti i troška, prikazani su na slici 22 i kasnije su detaljnije prokomentarisani.



SLIKA 22: GRAFOVI PRECIZNOSTI (LEVO) I GUBITKA (DESNO) VGG16 MODELA ZA KARAKTERISTIKU LOG MEL SPEKTROGRAM

4.4 Diskusija rezultata

CNN koja korisiti mel spektrogram bez augmentacije ima preciznost od 53% (Slika 17) . Pokazuje znake prekomernog obučavanja. Funkcija gubitka na skupu za obučavanje opada konstantno, dok na skupu za testiranje više stagnira. To znači da model dobro uči na skupu za obučavanje, ali ne generalizuje dobro na nove podatke. Takođe, preciznost na skupu za obuku je veća od preciznosti na skupu za testiranje.

CNN koji koristi mel spektrogram sa augmentacijom (Slika 18) pokazuje neznatno bolje performance u odnosu na model bez augmentacije, preciznost je 56%. Gubitak je niži, a preciznost je viša u poređenju sa prethodnim modelom. Funkcija gubitka na podacima za obučavanje se smanjuje brže od funkcije gubitka na podacima za testiranje, što znači da se model previše

prilagođava trening skupu. Preciznost na test podacima je bliska preciznosti na podacima za obučavanje, što znači da se model manje više prilagođava trening skupu u odnosu na prethodni model. Overfitovanje je manje izraženo što pokazuje da je augmentacija efikasna u poboljšanju generalizacije.

CNN koji koristi MFCC bez augmentacije ima preciznost od 62% (Slika 19) . Precizost modela na skupu za obučavanje raste tokom obuke, dok tačnost na skupu za testiranje pokazuje blagi rast sa nekim oscilacijama. Ne primećuje se značajan jaz između tačnosti na skupu za obučavanje i skupu za testiranje, što ukazuje da model nije previše overfitovan. Funkcija gubitka na skupu za obučavanje i skupu za testiranje opada tokom obuke, što je poželjno. Ne primećuje se ni značajan jaz između gubitka na skupu za obučavanje i skupu za testiranje, što ukazuje da model nije preobučen.

CNN koji koristi MFCC sa augmentacijom ima preciznost od 64% (Slika 20) . Model pokazuje stabilan rast tačnosti tokom obuke. Mali je jaz između funkcije preciznosti na skupu za treniranje i obuku što ukazuje da model dobro generalizuje nad novim podacima. Gubitak modela se smanjuje tokom obuke, što ukazuje na to da se model poboljšava u predikciji.

Na svim graficima se može videti da se gubitak modela smanjuje tokom treninga, kako za skup podataka za obuku tako i za test skup podatka, kao i da se funkcija preciznosti povećava nad podacima za obuku i testiranje. Međutim za neke modele se može primetiti da se funkcija gubitka nad podacima za obučavanje brže smanjuje nego nad podacima za testiranje što ukazuje na prekomerno preobučavanje.

Preciznost nad podacima za obučavanje za LSTM (Slika 21) počinje oko 0.6 i postepeno raste, dostižući oko 0.85 na kraju treninga. Ovo sugeriše da model efikasno uči šablone iz trening podataka. Preciznost na validacionom skupu prati sličan trend, počinje niže oko 0.55 i takođe raste do oko 0.8 na kraju. Međutim, postoji razlika između krivih preciznosti za trening i validaciju, pri čemu je preciznost na trening skupu viša. Ovo može ukazivati na određeni stepen preobučavanja, gde model previše dobro uči trening podatke i ne generalizuje dovoljno dobro na nove, neviđene uzorke. Postoje određeni skokovi i fluktacije u krivama funkcije gubitka za obučavanje i testiranje što može sugerisati da model ima poteškoća sa stabilnom konvergencijom.

VGG16 model je postigao tačnost od 70% (Slika 22), međutim, uočava se značajna razlika između krivih tačnosti za trening i validaciju, pri čemu je tačnost na trening skupu viša. Ova razlika može ukazivati na moguće preobučavanje modela, što znači da model previše dobro uči specifičnosti trening podataka, ali ne generalizuje dovoljno dobro na nove, neviđene uzorke. Gubitak na trening skupu počinje sa vrednošću od oko 1.5 i postepeno opada, dostižući oko 0.60 na kraju, što se poklapa sa povećanjem tačnosti na trening skupu. Slično, gubitak na validacionom skupu pokazuje opadajući trend, ali ostaje viši od gubitka na trening skupu, što ponovo ukazuje na prisustvo preprilagođavanja.

Sve u svemu modeli napreduju u učenju, ali postoji prostor za poboljšanje u smislu smanjenja preprilagođavanja i stabilizacije procesa treninga. Potencijalni sledeći koraci mogli bi uključivati:

- Istraživanje drugih tehnika regularizacije za poboljšanje generalizacije
- Podešavanje hiperparametara poput stope učenja, veličine batch-a itd
- Razmatranje složenije arhitekture modela, ako podaci to mogu podržati
- Proširivanje skupa podataka

5. Zaključak

Cilj ovog rada bio je da se istraže metode za prepoznavanje emocija iz govora uz implementaciju dubokih neuronskih mreža, sa fokusom na modele kao što su CNN, LSTM i VGG16. Koristeći RAVDESS skup podataka obrađeni su podaci, izvedena je ekstrakcija karakteristika i primenjeni su na spomenute modele kako bi se klasifikovalo 8 emocija.

U radu je objašnjena arhitektura spomenutih modela i proces izvlačenja značajnih karakteristika iz govornog signala. Kroz implementaciju i eksperimente dobijeni su zadovoljavajući rezultati za dati skup podataka.

Inicijalni modeli konvolucionih neuronskih mreža postigli su tačnosti od 53% za karakteristiku govora- mel spektrogram i 62% za MFCC. Upotrebom tehnika augmentacije podataka minimalno su poboljšeni rezultati, postižući preciznosti od 55% za mel spektrogram i 65% za MFCC. Za LSTM model dobijeni su rezultati od 68% koristeći MFCC sa augmentacijom. Najveća tačnost od 70%, se dobila primenom već treniranog modela VGG16.

Jedno ograničenje je bilo korišćenje male količine podataka, jer skup podataka sadrži 1.440 fajlova. U budućnosti mogu da se iskoriste dodatni skupovi podataka kao što su SAVEE, EMO DB, CREMA kao i da se probaju drugačiji parametri modela.

Dalji korak bio bi razvoj modela koji uzima u obzir i izraze lica i fiziološke signale(npr. otkucaje srca).

Literatura

[1] Milana M. Milošević (2020.) Identifikacija Govornika U Uslovima Emotivnog Govora, doktorska disertacija

[2] Antonina Stefanowska, Slawomir K. Zielinski, „Speech Emotion Recognition Using a Multi-Time-Scale Approach to Feature Aggregation and an Ensemble of SVM Classifiers“, doi: 10.24425/aoa.2024.148784(2024)

[3] B. Schuller, G. Rigoll i M. Lang, „Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture,“ u IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP’04, Montreal, 2004.

[4] Pyrovolakis, Konstantinos & Tzouveli, Paraskevi & Stamou, Giorgos. (2022). Multi-Modal Song Mood Detection with Deep Learning. Sensors. 22. 1065. 10.3390/s22031065.

[5] Ibrahim Patel (2010.) Speech Recognition Using HMM with MFCC-An Analysis Using Frequency Specral Decomposion Technique, Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.2, December 2010

[6] Oppenheim, A. V., & Schafer, R. W. (2010). *Discrete-Time Signal Processing*. Pearson Education

[7] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors (Switzerland), 21(4), 1–27. https://doi.org/10.3390/s21041249

[8] Skripta "Mašinsko učenje" profesora Mladena Nikolića

[9] "Introduction to RNN and LSTM(Part-1)", 2020., https://mc.ai/introduction-to-rnnand-lstmpart-1-2

[10] https://www.geeksforgeeks.org/vgg-16-cnn-model/

[11] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[12] Valerio Velardo URL: https://www.youtube.com/watch?v=bm1cQfb_pLA&ab_channel=ValerioVelardo-TheSoundofAI

[13] https://www.kaggle.com/code/blurredmachine/vggnet-16-architecture-a-complete-guide