SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I

INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Sveučilišni diplomski studij

Prepoznavanje glasa osobe

Lazar Volarev

Osijek, 2024.

##### SADRŽAJ

[1. UVOD 1](#_Toc168673299)

[2. NAČELNI PRINCIP RJEŠAVANJA PROBLEMA PREPOZNAVANJA GLASA 2](#_Toc168673300)

[3. PRISTUPI RJEŠAVANJU PROBLEMA PREPOZNAVANJA GLASA 3](#_Toc168673301)

[4. ODABRANI PRISTUP U OVOM PROJEKTU 4](#_Toc168673302)

[5. OPIS KORIŠTENIH SKUPOVA PODATAKA 5](#_Toc168673303)

[6. OPIS KORIŠTENIH METODA 6](#_Toc168673304)

[7. EVALUACIJA IZGRAĐENOG MODELA 7](#_Toc168673305)

[8. DEMONSTRACIJA RADA 8](#_Toc168673306)

[9. ZAKLJUČAK 9](#_Toc168673307)

# UVOD

Prepoznavanje glasa predstavlja proces identifikacije pojedinca na temelju njihovog glasa, što se često koristi u biometrijskim sustavima ili sigurnosnim aplikacijama. Ovaj projekt je imao za cilj izraditi rješenje temeljeno na dubokim neuronskim mrežama koje na temelju snimke glasa osobe prepoznaje o kojoj se osobi radi. Korišteni su javno dostupni skupovi podataka za treniranje i testiranje modela, te je razvijena aplikacija u Streamlitu koja omogućava upload zvučnog zapisa i klasifikaciju.

# NAČELNI PRINCIP RJEŠAVANJA PROBLEMA PREPOZNAVANJA GLASA

Prepoznavanje glasa, poznato i kao identifikacija govornika, proces je kojim se analizom govornog signala identificira pojedinac. Ovaj zadatak može se podijeliti u nekoliko ključnih koraka:

1. **Prikupljanje podataka:**
   * Prikupljanje velikog skupa podataka koji sadrže govorne zapise različitih govornika.
   * Podaci moraju biti raznoliki, uključujući varijacije u tonu, naglasku, brzini govora i uvjetima snimanja.
2. **Predobrada podataka:**
   * Uklanjanje šumova i neželjenih smetnji iz audio zapisa.
   * Normalizacija audio signala kako bi se osigurala konzistentnost podataka.
   * Segmentacija audio snimaka u manje dijelove za bolju obradu.
3. **Ekstrakcija značajki:**
   * Izvlačenje relevantnih značajki iz audio signala koje mogu pomoći u razlikovanju govornika.
   * Uobičajene značajke uključuju mel-frekvencijske kepstralne koeficijente (MFCC), spektralne značajke, prosodijske značajke i druge.
4. **Treniranje modela:**
   * Korištenje algoritama strojnog učenja i dubokih neuronskih mreža za treniranje modela na temelju ekstraktiranih značajki.
   * Podjela podataka na skupove za treniranje, validaciju i testiranje.
5. **Evaluacija i validacija:**
   * Procjena performansi modela korištenjem različitih metrika kao što su točnost, preciznost, odziv i F1 score.
   * Optimizacija hiperparametara kako bi se postigla bolja generalizacija modela.

# PRISTUPI RJEŠAVANJU PROBLEMA PREPOZNAVANJA GLASA

Postoji nekoliko pristupa rješavanju problema prepoznavanja glasa:

1. **Tradicionalni pristupi:**
   * **GMM-UBM (Gaussian Mixture Model - Universal Background Model):**
     + Koristi Gaussove mješavine za modeliranje distribucije značajki govora različitih govornika.
   * **HMM (Hidden Markov Models):**
     + Koristi skriveni Markovljev model za modeliranje sekvencijalnih podataka, poput govora.
2. **Strojno učenje:**
   * **Support Vector Machines (SVM):**
     + Koristi se za klasifikaciju govornika na temelju značajki izvučenih iz audio signala.
   * **K-Nearest Neighbors (KNN):**
     + Jednostavan algoritam koji klasificira govornike na temelju sličnosti s najbližim susjedima u prostoru značajki.
3. **Duboko učenje:**
   * **Konvolucijske Neuronske Mreže (CNN):**
     + Korištene za automatsku ekstrakciju značajki iz audio spektra.
   * **Rekurentne Neuronske Mreže (RNN) i Long Short-Term Memory (LSTM):**
     + Korištene za modeliranje vremenskih ovisnosti u govornim signalima.
   * **Speaker Embeddings (npr. x-vectors, d-vectors):**
     + Korištenje dubokih neuronskih mreža za stvaranje kompaktnog vektorskog prikaza govornika koji se zatim koristi za klasifikaciju.

# ODABRANI PRISTUP U OVOM PROJEKTU

U okviru ovog projektnog zadatka odabran je pristup temeljen na dubokim neuronskim mrežama zbog njihove sposobnosti da automatski nauče složene reprezentacije iz sirovih audio podataka. Konkretno, korištena je kombinacija konvolucijskih i rekurentnih neuronskih mreža:

1. **Konvolucijske neuronske mreže (CNN):**
   * Korištene su za ekstrakciju značajki iz spektrograma audio signala. CNN može automatski naučiti relevantne značajke bez potrebe za ručnim odabirom.
2. **Rekurentne neuronske mreže (RNN) i Long Short-Term Memory (LSTM):**
   * Korištene su za modeliranje sekvencijalne prirode govora. LSTM mreže su posebno korisne jer mogu zadržati dugoročne ovisnosti unutar govornog signala.
3. **Treniranje i evaluacija:**
   * Modeli su trenirani na velikom skupu podataka preuzetom iz LibriVox govorne baze podataka. Skup podataka podijeljen je na treniranje i testiranje s ukupno 360 sati čistih govora za treniranje i dodatnih 100 sati za testiranje.
   * Evaluacija je obavljena korištenjem različitih metrika kao što su točnost, preciznost, odziv i F1 score, te je model postigao visoku točnost prepoznavanja.

Odabrani pristup omogućava visoku točnost prepoznavanja glasa i može se lako skalirati na veće skupove podataka i različite jezike ili dijalekte.

# OPIS KORIŠTENIH SKUPOVA PODATAKA

Skup podataka korišten u ovom projektu preuzet je iz LibriVox govorne baze podataka. Podaci uključuju audio snimke govornika različitih spolova, koje su kategorizirane i označene za treniranje i testiranje modela. Eksplorativna analiza podataka uključivala je pregled distribucije trajanja snimaka po govornicima, spolovima i skupovima. Pred obrada podataka obuhvatila je uklanjanje šumova, normalizaciju audio signala i segmentaciju snimaka u manje dijelove za bolju obradu modela.

# OPIS KORIŠTENIH METODA

Za prepoznavanje glasa korišteni su modeli temeljeni na dubokim neuronskim mrežama, posebno konvolucijske i rekurentne neuronske mreže. Modeli su implementirani koristeći PyTorch biblioteku. Konvolucijske mreže korištene su za ekstrakciju značajki iz audio signala, dok su rekurentne mreže korištene za analizu sekvencijalnih podataka. Model je treniran korištenjem unaprijed podijeljenih skupova podataka uz optimizaciju hiperparametara i validaciju.  
Tokom treniranja, korištena je funkcija gubitka koja mjeri razliku između predviđenih i stvarnih oznaka, te su praćenje metrike kao što su točnost i preciznost.

# EVALUACIJA IZGRAĐENOG MODELA

Evaluacija modela obavljena je korištenjem različitih metrika kao što su točnost, preciznost, odziv i F1 score. Na testnom skupu podataka, model je postigao visoku točnost prepoznavanja, dok je matrica zabune pokazala učinkovitost modela u razlikovanju različitih govornika. Pored toga, izvršena je evaluacija performansi sustava u stvarnim uvjetima putem subjektivnih ocjena korisnika, gdje su korisnici ocjenjivali točnost prepoznavanja glasa na temelju stvarnih audio snimaka.

# DEMONSTRACIJA RADA

Aplikacija razvijena u Streamlit-u omogućuje korisnicima jednostavno učitavanje audio zapisa i pokretanje klasifikacije glasa. Sučelje aplikacije omogućuje korisnicima da prenesu zvučni zapis, a zatim klikom na gumb pokrenu proces prepoznavanja glasa. Rezultati klasifikacije prikazani su u obliku identificiranog govornika zajedno s vjerojatnošću prepoznavanja.

A screenshot of a black screen

Description automatically generated

Slika 1: Izgled sučelja aplikacije

Nakon učitavanja zvučnog zapisa, aplikacija prikazuje predviđenog govornika. Korisnici mogu vidjeti ime prepoznatog govornika.

A screenshot of a computer

Description automatically generated

Slika 2: Način prikaza rezultata

# ZAKLJUČAK

U okviru ovog projekta razvijen je sustav za prepoznavanje glasa koji koristi duboke neuronske mreže za identifikaciju govornika. Model je postigao visoku točnost na testnim podacima i uspješno je integriran u korisničku aplikaciju razvijenu u Streamlit-u. Moguća unaprjeđenja uključuju dodatnu optimizaciju modela, proširenje skupa podataka za bolju generalizaciju i implementaciju dodatnih sigurnosnih mjera u aplikaciji.