Röstbaserad Musikrekommendationer

En integration av maskininlärning och Spotify API för att identifiera känslor och skapa musikrekommendationer



Parsan Amani

EC Utbildning

2024-10

# Abstract

This project aimed to develop an application that uses voice-based emotion recognition to recommend music based on the user's mood. By integrating a machine learning model with the Spotify API, the system analyzed voice data to identify emotions such as happiness, anger, or neutrality. A 1D Convolutional Neural Network (CNN) was trained on MFCC features extracted from audio data, achieving a test accuracy of 93%. Data augmentation techniques, such as adding noise, were used to improve the model’s robustness. The final application successfully delivered personalized music recommendations based on the predicted emotional state of the user.

# Förkortningar och Begrepp

# CNN: Convolutional Neural Network

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[Abstract 2](#_Toc181288364)

[Förkortningar och Begrepp 3](#_Toc181288365)

[CNN: Convolutional Neural Network 3](#_Toc181288366)

[1 Inledning 1](#_Toc181288367)

[1.1 Problemdefinition 1](#_Toc181288368)

[2 Teori 2](#_Toc181288369)

[2.1 Neurala Nätverk 2](#_Toc181288370)

[2.1.1 Regularisering i CNN 2](#_Toc181288371)

[2.1.2 Välja Hyperparametrar 2](#_Toc181288372)

[2.2 MFCC (Mel Frequency Cepstral Coefficients) 2](#_Toc181288373)

[3 Metod 3](#_Toc181288374)

[3.1 Datainsamling 3](#_Toc181288375)

[3.2 Agil arbetsmetodik 4](#_Toc181288376)

[4 Resultat och Diskussion 5](#_Toc181288377)

[5 Slutsatser 6](#_Toc181288378)

[6 Självutvärdering 7](#_Toc181288379)

[Appendix A 8](#_Toc181288380)

[Källförteckning 9](#_Toc181288381)

# Inledning

Projektet syftade till att skapa en applikation som kunde analysera användarens röst för att identifiera känslor och därefter rekommendera musik via Spotify API. Den tekniska bakgrunden bygger på maskininlärning, där en modell tränades för att klassificera känslor utifrån röstdata. Relevansen för detta arbete ligger i den ökande efterfrågan på personanpassade musikupplevelser som kan förstärka användarens känslomässiga tillstånd i realtid.

Syftet med denna rapport är att beskriva processen för utveckling av känsloigenkänning via röst samt integreringen med Spotify API, för att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Hur tränades modellen för att identifiera känslor i ljuddata?
2. Hur kan modellen integreras med en applikation för att ge musikrekommendationer i realtid?

## Problemdefinition

I dagens digitala landskap blir det allt viktigare att skapa användaranpassade upplevelser, särskilt inom musikstreaming. Många nuvarande rekommendationssystem bygger på tidigare lyssningshistorik och individuella preferenser, men det saknas lösningar som kan anpassas dynamiskt efter användarens känslomässiga tillstånd i realtid. Det här projektet vill fylla den luckan genom att använda känsloigenkänning från röstdata i en musikrekommendationsapp. Utmaningen ligger i att kunna identifiera känslor från rösten med precision och sedan para ihop dessa känslor med lämplig musik.

# Teori

Projektet använde sig av maskininlärning för att utveckla en modell som kunde identifiera känslor från röstdata. Den metod som valdes var en **1D Convolutional Neural Network (CNN)**, som är särskilt lämplig för att bearbeta sekventiell data som ljud. CNN-modeller har visat sig vara effektiva för att identifiera mönster i komplexa dataset och har använts framgångsrikt inom områden som bildigenkänning och talanalys.

För att extrahera meningsfulla funktioner från ljudet användes **MFCC (Mel Frequency Cepstral Coefficients)**, en vanligt använd teknik inom taligenkänning. MFCC omvandlar ljudsignaler till en representation som bättre reflekterar hur människor uppfattar ljudfrekvenser, vilket gör det möjligt för modellen att identifiera känslor utifrån röstens ton och intonation.

För att förbättra modellens generaliseringsförmåga användes även **dataförstärkning** genom att lägga till brus till ljudfilerna. Detta bidrog till att modellen blev mer robust mot variationer i ljudkvalitet och olika inspelningsmiljöer.

## Neurala Nätverk

I detta projekt användes ett **1D Convolutional Neural Network (CNN)**, vilket är ett kraftfullt verktyg för att analysera sekventiell data som ljud. Ett CNN består av flera lager som hjälper till att extrahera och identifiera viktiga mönster i indata, såsom känslor som kan uttryckas i rösten. Modellen är speciellt lämpad för uppgifter där det är viktigt att fånga lokala mönster i data.

### Regularisering i CNN

För att förhindra att modellen överanpassar sig till träningsdata användes **L2-regularisering** och **Dropout**. Dessa tekniker reducerar modellens beroende av enskilda neuroner och hjälper till att generalisera till nya, osedda data.

### Välja Hyperparametrar

Under modellutvecklingen justeras hyperparametrar som antalet filter storleken på konvolutionskärnorna och inlärningshastigheten. Dessa val av parametrar påverkade modellens förmåga att identifiera mönster i ljudet utan att riskera överanpassning

## MFCC (Mel Frequency Cepstral Coefficients)

För att extrahera meningsfulla funktioner från ljudet användes **MFCC**-tekniken. MFCC omvandlar ljudsignaler till en representation som bättre reflekterar hur människor uppfattar ljudfrekvenser. Denna teknik är vanligt förekommande inom tal- och känsloigenkänning.

# Metod

Projektet använde en metod där ljudfiler analyseras för att bedöma användarens humör och därmed rekommendera passande Spotify-spellistor. Genom att använda Spotify API hanteras autentisering och spellistor hämtas baserat på användarens känslostämning. För ljudanalysen utnyttjas Librosa för att extrahera MFCC (Mel-Frequency Cepstral Coefficients), vilket är viktiga egenskaper för att identifiera tonhöjd och koppla detta till emotionella tillstånd. En TensorFlow-modell används sedan för att analysera dessa egenskaper och förutsäga en humörkategori, vilket möjliggör att rekommenderad musik bättre speglar användarens känslotillstånd i realtid.

I projektet bidrog Muhannad med utvecklingen av API som utför ljudanalysen och integrerar med Spotify för att ge personliga musikrekommendationer. David utvecklade en desktopapplikation som kopplar samman API och Spotify för att skapa en intuitiv och användarvänlig upplevelse. Ming anslöt sig till projektet en vecka efter start och arbetade med både applikationen och modellen för att säkerställa att alla delar fungerade tillsammans. Tillsammans utformade teamet en helhetslösning som både analyserar användarens humör och levererar en relevant spellista i realtid, vilket ger en personlig och engagerande musikupplevelse.

## Datainsamling

För detta projekt användes fyra olika dataset som alla fokuserade på att identifiera känslor genom röstanalys. Data hämtades från Kaggle och andra källor online. Ett av dessa dataset innehöll röstinspelningar på både engelska och mandarin, men endast den engelska delen användes för att säkerställa enhetlighet i språkanalysen.

De fyra dataset som användes inkluderade:

1. **Kannada Emotions Dataset**
2. **RAVDESS Dataset**
3. **Toronto Emotional Speech Set (TESS)**
4. **Emotional Speech Dataset (ESD)**

Varje dataset kategoriserade ljudfilerna efter specifika känslomässiga tillstånd som "angry" (arg), "disgust" , "fear" , "happy" , "neutral", "sad" och "surprised" , enligt den visade figurens struktur. Efter att datan hämtats genomgick den en rengöringsprocess där brus och oönskade ljudsekvenser togs bort. Endast klara och tydliga ljudfiler, med starka känslomässiga uttryck, togs med i datasetet för både träning och testning.

## Agil arbetsmetodik

Enligt de agila principerna strukturerades projektet i korta, hanterbara iterationer, kallade sprintar, där varje komponent levererades och utvärderades löpande. Teamet bestod av fyra personer med olika ansvarsområden, men vi arbetade tätt tillsammans och hade regelbundna möten var tredje till fjärde dag för att gå igenom framsteg och anpassa planen vid behov.

De huvudsakliga agila principerna vi följde var:

* **Tidig och kontinuerlig leverans**: Genom att arbeta med mindre projektmoduler, som modellering och integration, kunde vi testa och bekräfta att varje del fungerade innan vi fortsatte.
* **Flexibilitet inför förändrade krav**: Under projektets gång behövde vi flera gånger anpassa både modellen och integrationen, särskilt när vi stötte på tekniska utmaningar som integrationen med Spotify API.
* **Självorganiserande team**: Varje medlem hade ansvar för sitt specifika område – modellering, API-integration eller applikationsutveckling – men vi hjälptes åt för att lösa problem och hitta gemensamma lösningar.

Projektarbetet delades in i tre huvuddelar där varje medlem ansvarade för sitt område:

* David ansvarade för applikationen och samarbetade kring integrationen.
* Muhannad fokuserade på pipelinen och bidrog även till integrationsdelen.
* Parsan arbetade med modellen och dataset samt med integrationen.
* Ming arbetade med både applikationen och modellen och deltog också i integrationsarbetet.

# Resultat och Diskussion

Modellen som tränades med hjälp av MFCC-funktioner från röstdata presterade med följande resultat:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Modell** |  |  | | --- | |  | | | **Test Accuracy** | | --- |  |  | | --- | |  | |
| CNN-modell (med regularisering) | 93% |

Tabell 1: Resultat av modellens testnoggrannhet efter träning och validering.

Den slutgiltiga modellen nådde en **testnoggrannhet** på 93%, vilket indikerar att modellen är mycket kapabel att korrekt identifiera känslor baserat på röstdata. Det är dock viktigt att notera att vissa känslor, såsom "neutral" och "surprised", visade sig vara mer svåridentifierade på grund av deras akustiska likheter.

Resultaten visar att användningen av **dataförstärkning** (som att lägga till brus) förbättrade modellens prestanda genom att göra den mer robust mot variationer i inspelningsmiljö och ljudkvalitet. Genom att använda en **1D CNN-modell** med **L2-regularisering** och **Dropout** lyckades vi minska överanpassning och förbättra generalisering till nya data.

En begränsning i modellen var dock dess förmåga att exakt identifiera vissa känslor, särskilt när ljudkvaliteten varierade kraftigt eller när känslorna var liknande i tonhöjd och intensitet.

Vid en framtida utveckling av systemet skulle vi kunna dra nytta av ett större dataset samt mer avancerade förbehandlingsmetoder för att ytterligare förbättra modellens noggrannhet.

# Slutsatser

Detta projekt visade att det är möjligt att använda röstbaserad känsloigenkänning för att generera personliga musikrekommendationer via integration med Spotify API. Genom att använda en **1D CNN-modell** tränad på MFCC-funktioner, tillsammans med dataförstärkning genom brus, uppnåddes en hög testnoggrannhet på 93%, vilket bekräftar modellens förmåga att korrekt identifiera känslor i röstdata.

1. **Hur tränades modellen för att identifiera känslor i ljuddata?**  
   Modellen tränades med hjälp av MFCC-funktioner samt **dataförstärkning** där brus lades till ljudet för att simulera olika inspelningsmiljöer och göra modellen mer robust. Genom användning av L2-regularisering och dropout-tekniker undveks överanpassning och modellen uppnådde god generalisering.
2. **Hur kan modellen integreras med en applikation för att ge musikrekommendationer i realtid?**  
   Genom integrationen med Spotify API kan den tränade modellen användas i en desktopapplikation för att analysera användarens röst och direkt generera musikrekommendationer baserade på det identifierade känsloläget. Detta möjliggör en dynamisk och personlig upplevelse för användaren.
3. **Framtida arbete**

Trots goda resultat finns det möjligheter till förbättringar, såsom att inkludera ett större dataset eller använda mer avancerade förbehandlingsmetoder. Att inkludera fler typer av dataförstärkning samt att optimera modellen för fler känslonyanser och språkliga variationer skulle ytterligare förbättra precisionen i rekommendationerna.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Jag stötte på flera utmaningar under projektet, från datainsamling till modellering och utvärdering. I början var det problem med att hantera och strukturera datamängden från olika källor, där vissa dataset saknade vissa känslokategorier, vilket ledde till obalans i datasetet. Detta kan ha påverkat modellen eftersom vissa kategorier som "surprised" och "calm" hade betydligt fler exempel än andra, vilket kan ha gjort att modellen överanpassade till dessa känslor.

1. Vilket betyg du anser att du skall ha och varför.

Jag anser att jag förtjänar ett VG betyg för detta projekt eftersom jag har visat både uthållighet och god problemlösningsförmåga. Jag hanterade olika tekniska utmaningar och arbetade aktivt med att förbättra modellen, trots att jag stötte på obalanser i datasetet och kompatibilitetsproblem med olika bibliotek. Under projektet har jag också lärt mig mycket om dataaugmentation och hur denna teknik kan användas för att stärka modellen och kompensera för databegränsningar.

1. Något du vill lyfta fram till Antonio?

Inget särskilt jag vill lyfta fram just nu, tack för möjligheten att arbeta med detta projekt.

# Appendix A

# Källförteckning

<https://www.kaggle.com/datasets/pragyanaianddsschool/kannada-emotions>

<https://www.kaggle.com/datasets/descipro/ravedess-dataset>

<https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess>

<https://hltsingapore.github.io/ESD/download.html>