Clasificarea Limbii din Mesaje Audio

Giuglan Cătălin-Ionuț
Faculty of Automatic Control and Computer Science
UNSTPB
Bucharest, Romania
catalingiuglan@yahoo.com

***Abstract*—Acest proiect prezintă dezvoltarea unui sistem de detecție automată a limbii vorbite în înregistrări audio scurte, folosind un model Convolutional Recurrent Neural Network (CRNN) antrenat pe spectrograme Mel. Fișierele audio sunt preprocesate prin normalizare a volumului, eliminarea secvențelor inutile și ajustarea duratei la aceeași lungime, urmate de generarea spectrogramelor care sunt folosite ca intrare în model. Pentru a crește robustețea, sunt aplicate și mici variații ale vitezei de redare, simulând diferențele naturale dintre vorbitori. Sistemul suportă nouă limbi: germană, engleză, spaniolă, franceză, italiană, japoneză, portugheză, română și chineză, și este integrat într-o aplicație web Streamlit care afișează predicția și probabilitățile asociate. Rezultatele experimentale indică un model stabil, cu valori ridicate ale acurateții și scorului F1, reușind să clasifice corect în aproximativ 90% din cazuri.**

I. Introducere

Identificarea automată a limbii vorbite reprezintă o componentă esențială în multe aplicații moderne care procesează vorbirea. Sisteme precum asistenții virtuali, platformele de transcriere automată sau serviciile de analiză media trebuie adesea să determine mai întâi limba în care vorbește utilizatorul, pentru a selecta modelul de procesare adecvat. Acest proiect nu urmărește înțelegerea conținutului vorbit, ci doar identificarea limbii, folosind informații acustice precum ritmul vorbirii, pronunția și modul în care energia se distribuie pe frecvențe.

Acest proiect urmărește o abordare inspirată din lucrarea lui Singh et al. [1], unde sunt analizate beneficiile combinării rețelelor convoluționale cu rețele recurente pentru această sarcină. CNN-urile sunt eficiente în extragerea caracteristicilor locale din spectru, în timp ce RNN-urile sunt potrivite pentru a surprinde evoluția în timp a semnalului de vorbire. Împreună, aceste două componente permit modelului să identifice modele acustice specifice fiecărei limbi chiar și din intervale scurte de timp.

Pentru antrenare a fost folosit datasetul *VoxLingua107* [2], o colecție foarte mare care conține înregistrări din contexte reale, cu variații naturale de accent, intonație și calitate acustică. Pentru proiectul de față a fost selectat un subset alcătuit din nouă limbi: germană, engleză, spaniolă, franceză, italiană, japoneză, portugheză, română și chineză. Aceste limbi asigură o diversitate fonetică importantă, permițând evaluarea modelului atât pe limbi apropiate între ele, cât și pe limbi care sunt fonetic foarte diferite.

Scopul proiectului este dezvoltarea unui sistem capabil să identifice limba vorbită dintr-o mostră audio scurtă, într-un mod rapid, robust și ușor de utilizat, integrat într-o aplicație web accesibilă.

II. Setul de Date și Preprocesarea Audio

A. Exemple de Mel-Spectrograme

Pentru a ilustra diferențele acustice dintre limbi, figurile de mai jos prezintă două spectrograme Mel pentru înregistrări scurte în limba română și engleză. Spectrogramele evidențiază distribuția energiei în funcție de frecvență, o reprezentare foarte potrivită pentru analiza limbii, deoarece surprinde structuri fonetice și modele de pronunție dificil de observat în forma brută a semnalului audio.

Comparând cele două exemple, se observă diferențe vizibile în densitatea și variația benzilor de frecvență, derivate din particularitățile limbajului, precum ritmul vorbirii, durata medie a silabelor sau modul în care energia este concentrată în zone joase sau înalte ale spectrului. Aceste diferențe influențează forma generală a spectrogramelor Mel și justifică utilizarea lor ca intrare în model.

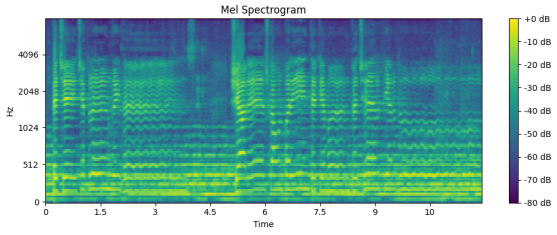

Figura 1. Spectrogramă Mel (limba română).

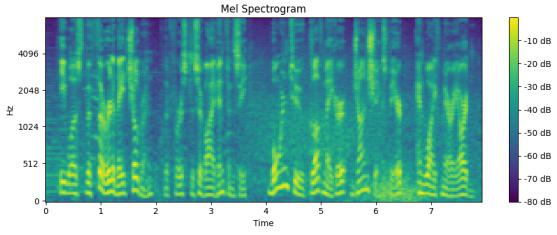

Figura 2. Spectrogramă Mel (limba engleză).

B. Preprocesare audio

Pentru ca modelul să poată învăța în mod eficient, toate mostrele audio sunt procesate printr-un pipeline standardizat.

Scopul acestor pași nu este doar uniformizarea datelor, ci și eliminarea elementelor inutile, reducerea zgomotului și creșterea rezistenței modelului la situații acustice variate.

Procesarea include:

- resampling la 16 kHz pentru a avea același format pentru toate fișierele;
- eliminarea secvențelor de ”liniște” (în care nu se aude nimic), astfel încât modelul să învețe din vorbire, nu din pauze;
- normalizare RMS, care uniformizează volumul înregistrărilor;
- tăiere sau completare la 3 secunde, astfel încât toate fișierele audio să aibă aceeași durată;
- generarea spectrogramelor Mel;
- conversia în decibeli pentru a pune în evidență structura spectrală.

Această etapă asigură faptul că modelul învață doar informațiile relevante, reducând diferențele artificiale dintre înregistrări (ex: volum, pauze, durată).

III. Procesul de Antrenare

Modelul CRNN a fost antrenat timp de 30 de epoci, cu o rată de învățare de mică, pentru a . Pentru a preveni blocarea într-un minim local, a fost folosit un scheduler *ReduceLROnPlateau*, iar pentru a evita suprantrenarea s-a aplicat *early stopping*.

Figura 3 arată evoluția funcției de pierdere și a preciziei. Loss-ul pe setul de antrenare scade rapid, de la aproximativ 1.62 la 0.08, ceea ce indică faptul că modelul învață progresiv să diferențieze mult mai bine cele nouă limbi.

În același timp, precizia pe setul de validare crește constant, pornind de la aproximativ 60% în primele epoci și depășind 92% în ultimele. Evoluția armonioasă dintre scăderea loss-ului și creșterea preciziei indică un model stabil și bine echilibrat.

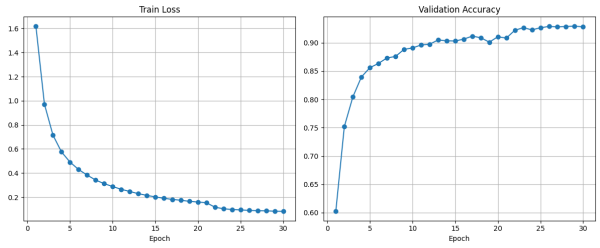

Figura 3. Evoluția metricilor.

A. Scor F1

Pentru a evalua performanța și în situații în care clasele pot fi dezechilibrate, a fost analizat și scorul F1 (Figura 4). Acesta rămâne constant ridicat, ceea ce confirmă faptul că modelul nu doar clasifică bine, ci și menține un echilibru între precizie și recall pentru majoritatea limbilor.

B. Matricea de Confuzie

Matricea de confuzie din Figura 5 oferă o perspectivă detaliată asupra modului în care modelul diferențiază limbile. Se observă că performanța este ridicată pentru toate clasele,

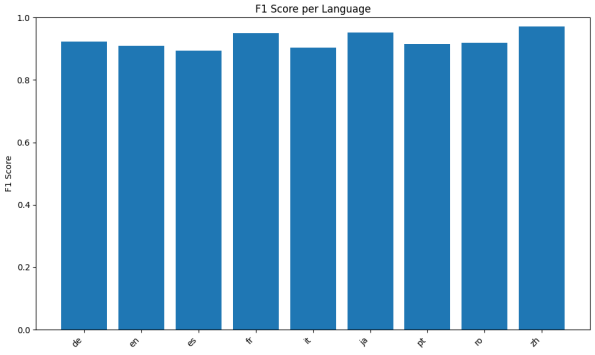

Figura 4. Evoluția scorului F1.

însă apar unele confuzii între limbile romanice (italiană, spaniolă, portugheză), ceea ce este de așteptat având în vedere asemănările fonetice și ritmice dintre acestea. Pe de altă parte, limbi precum japoneza sau chineza sunt aproape întotdeauna identificate corect, deoarece prezintă trăsături acustice distincte.

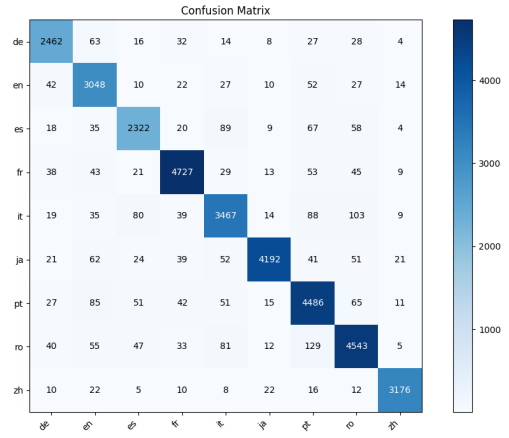

Figura 5. Matricea de confuzie pe setul de validare.

IV. Predicția Limbii și Interfața Web

A. Distribuția probabilităților

Pentru fiecare fișier audio analizat, modelul generează o distribuție de probabilități pentru toate cele nouă limbi (Figura 6). Aceasta permite utilizatorului să înțeleagă nu doar rezultatul final, ci și nivelul de încredere al modelului. Chiar și atunci când două limbi sunt apropiate, vizualizarea ajută la interpretarea corectă a deciziei.

B. Interfața aplicației Streamlit

Aplicația web dezvoltată în Streamlit (Figura 7) oferă o interfață accesibilă care permite încărcarea fișierelor audio, redarea acestora, efectuarea predicției și vizualizarea grafică a probabilităților. Aceasta face ca modelul să fie ușor de utilizat atât în scopuri demonstrative, cât și în contexte educaționale.

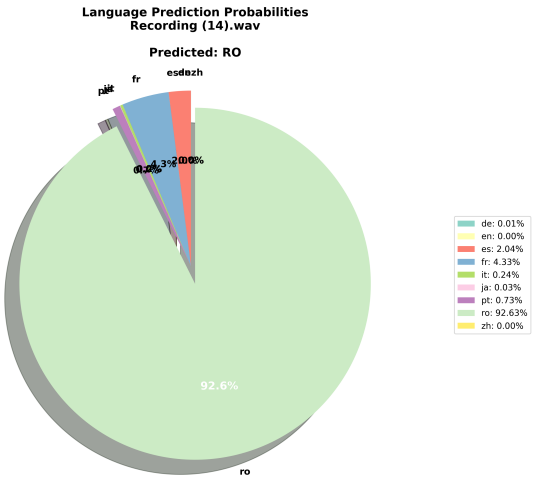

Figura 6. Distribuția probabilităților pentru un fișier audio.

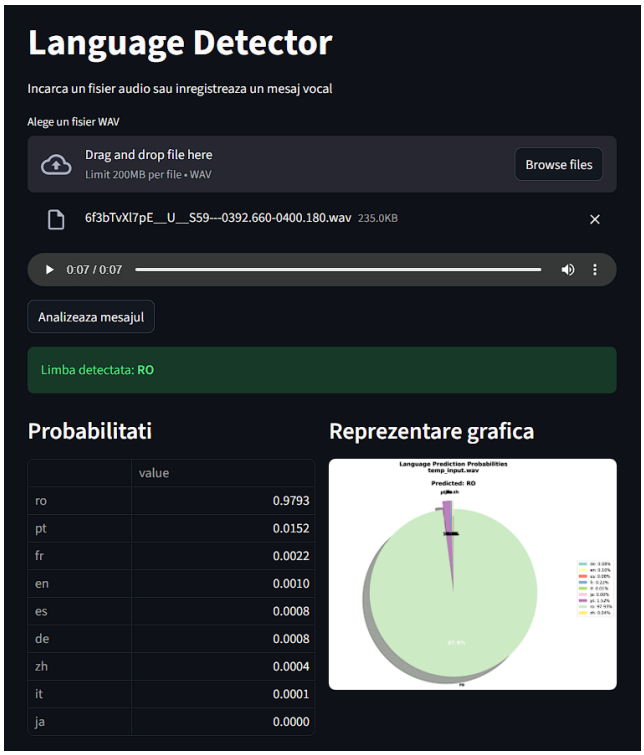

Figura 7. Interfața aplicației Streamlit.

V. Concluzii

Proiectul realizat arată că un model de tip CRNN poate identifica limba vorbită chiar și din mostre audio foarte scurte, de numai câteva secunde. Rezultatele obținute în timpul antrenării arată că modelul învață bine diferențele dintre cele nouă limbi folosite, iar scăderea constantă loss-ului și creșterea preciziei confirmă faptul că procesul de antrenare a decurs corect, fără să apară probleme precum overfitt.

Analizând performanța pe setul de validare, se observă că modelul reușește să facă predicții corecte în majoritatea cazurilor. Chiar și pentru limbi apropiate, cum ar fi italiana, spaniola sau portugheza, rezultatele sunt bune, iar confuziile sunt în general rezultate ale similarităților fonetice dintre aceste limbi. Scorul F1 ridicat și matricea de confuzie arată că modelul nu doar „ghicește” limba, ci are un comportament echilibrat pentru toate clasele.

Integrarea modelului într-o aplicație web dezvoltă partea practică a proiectului și îl face mult mai ușor de folosit. Utilizatorul poate încărca rapid un fișier audio, poate asculta înregistrarea și poate vedea într-un mod intuitiv atât predicția finală, cât și probabilitățile pentru fiecare limbă. Astfel, sistemul devine accesibil și util nu doar în scopuri academice, ci și pentru teste sau demonstrații.

În concluzie, proiectul demonstrează că un sistem relativ compact și ușor de antrenat poate oferi rezultate foarte bune în identificarea limbii vorbite. Pentru lucrări viitoare, sistemul poate fi îmbunătățit prin adăugarea mai multor limbi, antrenarea pe întregul dataset VoxLingua107 sau folosirea unor modele moderne pre-antrenate precum Wav2Vec2. Aceste îmbunătățiri ar putea crește și mai mult acuratețea și flexibilitatea sistemului.

Bibliografie

[1] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, M. Masud, “Spoken Language Identification Using Deep Learning,” *Computational Intelligence and Neuroscience*, 2021.

[2] T. Alumäe. VoxLingua107: A Large-Scale Multilingual Speech Dataset. https://cs.taltech.ee/staff/tanel.alumae/data/voxlingua107/