Spoken Language Classification from Audio Messages

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

***Abstract*—This project presents the development of an automatic spoken language detection system for short audio recordings, using a Convolutional Recurrent Neural Network (CRNN) trained on Mel spectrograms. The audio files are preprocessed through volume normalization, removal of irrelevant segments, and duration adjustment to a fixed length, followed by spectrogram generation used as model input. To improve robustness, small playback speed variations are applied, simulating natural speaker differences. The system supports nine languages: German, English, Spanish, French, Italian, Japanese, Portuguese, Romanian, and Chinese, and is integrated into a Streamlit web application that displays both the predicted language and the associated probabilities. Experimental results indicate a stable model with high accuracy and F1-score values, correctly classifying the spoken language in approximately 90% of cases.**

I. Introduction

Automatic spoken language identification is a key component in many modern speech-processing applications. Systems such as virtual assistants, automatic transcription platforms, or media analysis services often need to first determine the language spoken by the user in order to select the appropriate processing model. This project does not aim to understand the spoken content, but rather to identify the language itself, using acoustic information such as speech rhythm, pronunciation, and energy distribution across frequencies.

The approach adopted in this project is inspired by the work of Singh et al. [1], which analyzes the benefits of combining convolutional neural networks with recurrent neural networks for this task. CNNs are effective at extracting local features from spectral representations, while RNNs are well suited to capturing temporal evolution in speech signals. Together, these components allow the model to identify language-specific acoustic patterns even from short time intervals.

For training, the *VoxLingua107* dataset [2] was used, a large-scale collection containing recordings from real-world contexts, with natural variations in accent, intonation, and acoustic quality. For this project, a subset of nine languages was selected: German, English, Spanish, French, Italian, Japanese, Portuguese, Romanian, and Chinese. These languages provide significant phonetic diversity, allowing evaluation on both closely related languages and phonetically distinct ones.

The goal of this project is to develop a system capable of identifying the spoken language from a short audio sample in a fast, robust, and user-friendly manner, integrated into an accessible web application.

II. Dataset and Audio Preprocessing

A. Mel Spectrogram Examples

To illustrate acoustic differences between languages, the figures below present two Mel spectrograms corresponding to short recordings in Romanian and English. Spectrograms highlight the energy distribution across frequencies and represent a suitable input for language analysis, as they capture phonetic structures and pronunciation patterns that are difficult to observe in raw audio signals.

By comparing the two examples, visible differences can be observed in the density and variation of frequency bands, derived from language-specific characteristics such as speech rhythm, average syllable duration, or how energy is concentrated in lower or higher frequency regions. These differences influence the overall shape of Mel spectrograms and justify their use as model input.

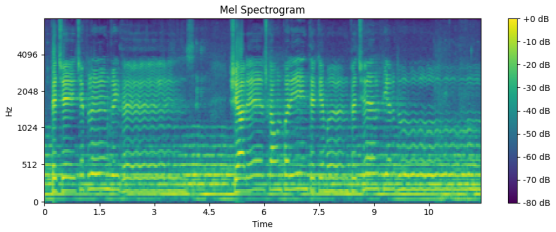

Figure 1. Mel spectrogram (Romanian language).

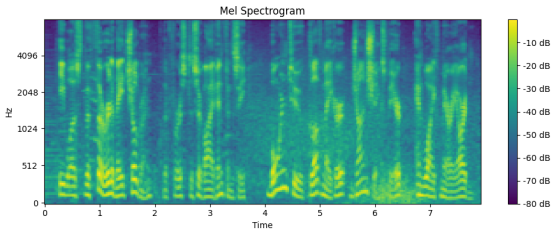

Figure 2. Mel spectrogram (English language).

B. Audio Preprocessing

In order for the model to learn efficiently, all audio samples are processed through a standardized pipeline. The purpose of these steps is not only to ensure data uniformity, but also to remove irrelevant information, reduce noise, and improve robustness under diverse acoustic conditions.

The preprocessing pipeline includes:

- resampling to 16 kHz to ensure a consistent format across all files;
- removal of silence segments, so the model learns from speech rather than pauses;
- RMS normalization to standardize recording volume;
- trimming or padding to 3 seconds, ensuring all audio files have the same duration;
- Mel spectrogram generation;
- conversion to decibel scale to emphasize spectral structure.

This stage ensures that the model learns only relevant information, reducing artificial differences between recordings such as volume, silence, or duration.

III. Training Process

The CRNN model was trained for 30 epochs with a low learning rate to ensure stable convergence. To prevent stagnation in local minima, a *ReduceLROnPlateau* scheduler was used, while *early stopping* was applied to avoid overfitting.

Figure 3 shows the evolution of the loss function and accuracy. The training loss decreases rapidly from approximately 1.62 to 0.08, indicating that the model progressively learns to better distinguish between the nine languages.

At the same time, validation accuracy increases steadily, starting from around 60% in the initial epochs and exceeding 92% in the final stages. The balanced evolution of decreasing loss and increasing accuracy indicates a stable and well-trained model.

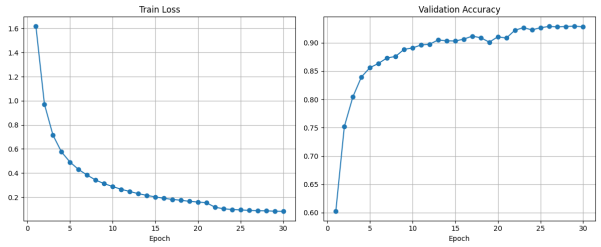

Figure 3. Training metrics evolution.

A. F1 Score

To evaluate performance in scenarios where class imbalance may exist, the F1 score was also analyzed (Figure 4). It remains consistently high, confirming that the model not only achieves high accuracy but also maintains a good balance between precision and recall across most languages.

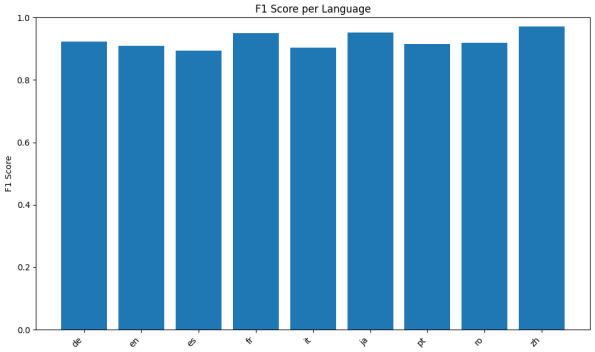

Figure 4. F1-score evolution.

B. Confusion Matrix

The confusion matrix in Figure 5 provides a detailed view of how the model distinguishes between languages. High performance is observed across all classes, although some confusion appears among Romance languages (Italian, Spanish, Portuguese), which is expected given their phonetic and rhythmic similarities. In contrast, languages such as Japanese and Chinese are almost always correctly identified due to their distinct acoustic characteristics.

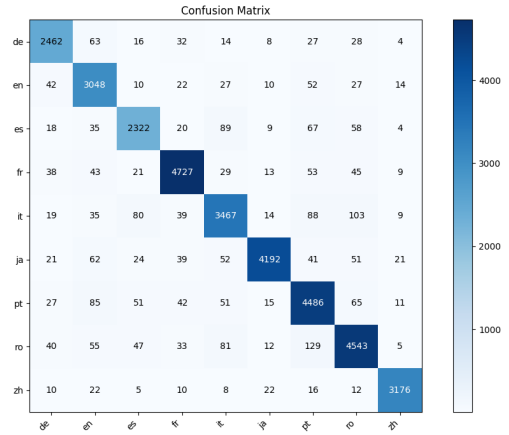

Figure 5. Confusion matrix on the validation set.

IV. Language Prediction and Web Interface

A. Probability Distribution

For each analyzed audio file, the model generates a probability distribution over all nine languages (Figure 6). This allows users to understand not only the final prediction but also the model's confidence. Even when languages are similar, this visualization helps interpret the decision process.

B. Streamlit Application Interface

The web application developed using Streamlit (Figure 7) provides an accessible interface that allows users to upload audio files, play them, perform predictions, and visualize probability distributions. This makes the model easy to use for both demonstration and educational purposes.

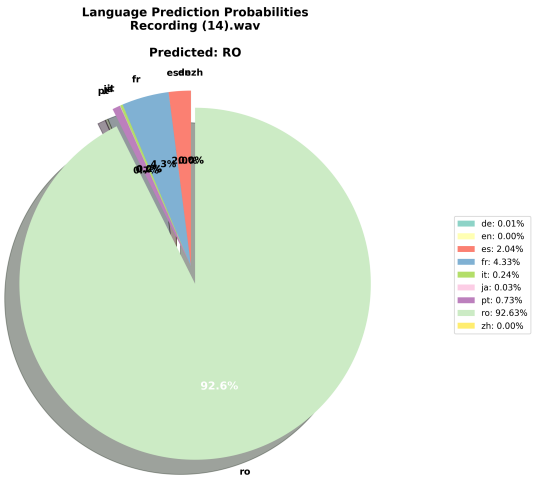

Figure 6. Probability distribution for an audio file.

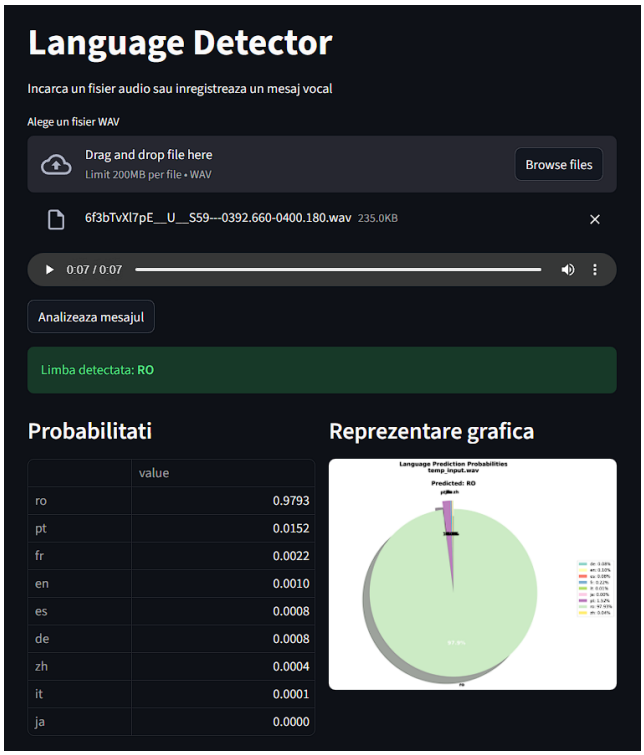

Figure 7. Streamlit application interface.

V. Conclusions

The developed project demonstrates that a CRNN-based model can accurately identify the spoken language even from very short audio samples of only a few seconds. The training results show that the model effectively learns the differences between the nine languages, and the consistent decrease in loss combined with increasing accuracy confirms a well-conducted training process without overfitting.

Evaluation on the validation set indicates that the model produces correct predictions in most cases. Even for closely related languages such as Italian, Spanish, and Portuguese, performance remains strong, with most confusions arising from inherent phonetic similarities. The high F1 score and confusion matrix analysis confirm that the model behaves in a balanced and reliable manner across all classes.

Integrating the model into a web application enhances the practical value of the project and makes it significantly easier to use. Users can quickly upload an audio file, listen to it, and view both the final prediction and the probability distribution in an intuitive format. This makes the system suitable not only for academic purposes, but also for testing and demonstrations.

In conclusion, this project shows that a relatively compact and efficient system can achieve very good performance in spoken language identification. Future work could include adding more languages, training on the full VoxLingua107 dataset, or leveraging modern pre-trained models such as Wav2Vec2. These improvements could further increase the accuracy and flexibility of the system.

References

[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/