**Spoken Language Identification using Deep Learning and PySpark**

**A report on**

**Spoken Language Identification using Deep Learning and PySpark**

**[CSE- 3263]**

Submitted By

**Boya Yashasvi Naidu 210962070**

**Akula Sanjay 210962016**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**MANIPAL INSTITUTE OF TECHNOLOGY,**

**MANIPAL ACADEMY OF HIGHER EDUCATION**

**APRIL 2024**

***Abstract—*** ***This project explores advanced techniques in spoken language identification using Convolutional Neural Networks (CNNs) with Mel Frequency Cepstral Coefficients (MFCC) and Filter Banks (FB) for feature extraction. By leveraging MFCC and FB features to capture effective speech representations, the study aims to enhance language detection accuracy and robustness. Integrating CNNs enables hierarchical feature learning from audio data, improving classification performance for languages like English, Spanish, and German. Through experimental validation on diverse datasets, this research demonstrates the effectiveness of combining CNNs with MFCC and FB features, particularly in noisy and challenging environments. This work contributes to advancing language identification technology with practical implications for communication systems and language-specific applications.***

1. INTRODUCTION

The increasing interconnectedness brought about by globalization highlights the need for effective cross-cultural communication. However, linguistic diversity poses a significant challenge to this communication. Language identification (LID) systems play a crucial role in addressing this challenge by automatically identifying spoken languages from speech samples, transcending factors like accent or pronunciation.

Humans have traditionally excelled in language identification, but with the complexity of global languages, machine learning (ML) and artificial intelligence (AI) offer scalable solutions. Modern approaches, including deep learning techniques like convolutional neural networks (CNNs), have revolutionized language identification by automatically learning relevant features from speech data.

In this project, a CNN-based LID system is developed to accurately identify languages such as English, Spanish, and German. Leveraging CNNs for their feature extraction capabilities, the project aims to enhance language identification accuracy, enabling applications like multilingual customer service and speech-to-text conversion.

By integrating advancements in CNN architectures with language identification, the project seeks to contribute to the field's evolution, ultimately enabling more seamless and efficient cross-language communication in diverse settings.

1. LITERATURE REVIEW

[1] The paper focuses on language identification from speech, proposing a methodology using machine learning and feature extraction techniques. It discusses the process of preprocessing data, extracting features such as Mel-frequency cepstral coefficients (MFCCs), and using algorithms like Linear Regression, Decision Tree, Random Forest, and Gradient Boosting for classification. The study includes the collection of a dataset comprising English, German, and Spanish audio samples, achieving high accuracy rates, with Random Forest yielding the best results (98% accuracy). The paper concludes by suggesting future improvements, including expanding the dataset, incorporating incremental machine learning, and adding more languages for classification.

[2] The paper discusses the development of a language identification (LID) system using speech samples from seven different Indian languages. The methodology involves pre-processing steps such as frame segmentation and noise removal, followed by feature extraction using Mel Frequency Cepstral Coefficients (MFCC) and formant frequencies. Clustering using K-Means is employed as a pre-processing step before classification with Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). Results indicate that LDA outperforms SVM in classification accuracy, with both MFCCs and formant frequencies contributing to improved accuracy when used together. The study suggests that speech features vary across languages, and further research could explore additional features and combinations to enhance system performance.

[3] The paper introduces a language identification method based on a dual attention mechanism, utilizing a Convolutional Block Attention Module (CBAM) and Bidirectional Gated Recurrent Unit (Bigru). The CBAM module enhances feature extraction by assigning weights to feature maps in channel and spatial dimensions, while the Bigru network extracts temporal sequence information. The model is evaluated on the Common voice dataset and AP17-OLR dataset, achieving high accuracy rates. Experiment results show that the proposed method outperforms other models in language identification tasks. Further research is suggested to explore languages with limited resources, such as small languages and dialects.

[4] The paper presents an Automatic Language Identification (LID) system based on Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction. The system is designed as a two-phase process: feature extraction from audio signals and classification using SVM and Decision Tree classifiers. The dataset includes recordings in English, Hindi, Tamil, and Telugu. Experimental results show an accuracy of 76% and 73% for SVM and Decision Tree classifiers respectively, demonstrating the effectiveness of the proposed methodology for language identification tasks.

[5] The paper focuses on developing a multilingual system for identifying Indian languages in both noisy and spoofing attack environments. The methodology involves using Mel Frequency Cepstral Coefficients (MFCC) and Constant Q Cepstral Coefficients (CQCC) as features for language identification and spoof detection. A CNN model is trained on concatenated MFCC and CQCC features, achieving an accuracy of 97% for language identification. The study uses a dataset containing synthetic voice samples for spoofing attacks and real voice samples for noisy environments, showing promising results for detecting spoofing attacks in Indian languages.

III. RESEARCH GAPS

Insufficient Handling of Noise: There is limited discussion on how noise robustness is addressed in the language recognition system. More detailed analysis and experimentation on noise robustness would be beneficial.

Scope of Study: The papers focus on resource-rich languages but does not cover resource-poor corpora such as small languages and dialects. This highlights a gap in the applicability and generalizability of the proposed method to a wider range of linguistic contexts.

Application Context: The papers focuses on the application of automatic language identification in specific scenarios, such as customer care centers and speech-to-text conversion. While these are important applications, they don’t explore the broader context of language identification, such as its use in multilingual societies or its implications for linguistic diversity and cultural preservation.

Limited Investigation of Spoofing Attacks: The primarily focus is on the Logical Access (LA) spoofing attack and does not explore other types of spoofing attacks that could be encountered in real-world scenarios.

IV. METHODOLOGY

A Convolutional Neural Network is trained after splitting the dataset in the ratio of 90:10.

The architecture of the model is explained in detail.

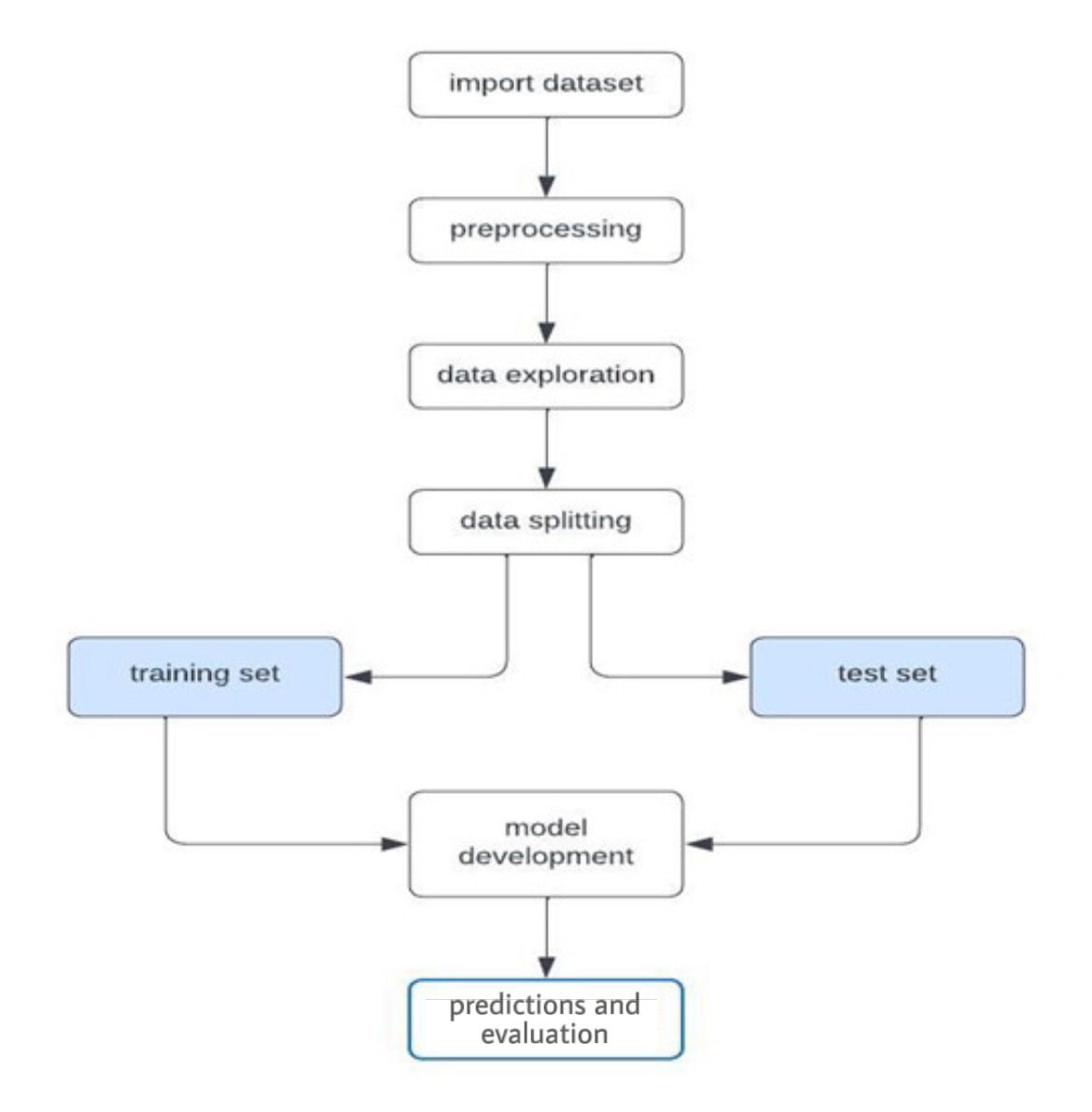


Fig 1: Block Diagram of the entire methodology

*A. Data Acquisition:*

The dataset utilized in this study comprises of audio files in three languages: English, Spanish and Deutsch in .flac format with 24380 files for each language.

*B. Preprocessing:*

The dataset underwent modifications involving essential feature updates. A Dataframe is created with the filename and its language of origin. This will be used to train the model.

*C. Data Exploration:*

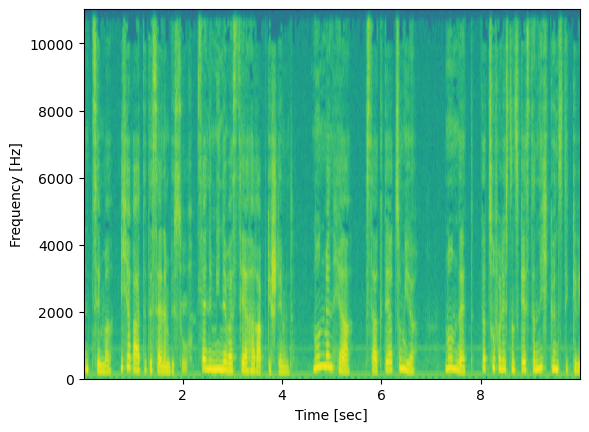


Fig 2: Spectrogram of a single file

A graph showing a sound wave

Description automatically generated

Fig 3: Traditional waveform (Amplitude vs time)

From the above two cases, we can conclude that the audio files are fairly consistent without much noise. Hence there isn’t a need to clean the audio files as such.

*D. Data Splitting:*

Divide the pre-processed dataset into training and testing sets using a 90:10 ratio. The training set (90%) will be used for model training, while the testing set (10%) will evaluate model performance.

*E. FB and MFCC:*

We implement a function to generate Filter Banks (FB) for audio signal processing, specifically for use in speech analysis, and subsequently convert them into Mel-Frequency Cepstral Coefficients (MFCC). Both FB and MFCC are critical features for various applications in speech processing and language detection tasks.

*Process of Generating FB and MFCC:*

1. Pre-Emphasis:

A pre-emphasis filter is applied to the input audio signal to amplify high frequencies.

This step helps in improving the signal-to-noise ratio and making the subsequent processing more effective.

2. Framing:

The pre-emphasized signal is divided into short frames using a sliding window technique (typically 25ms frame size with 10ms overlap).

Each frame is windowed using a Hamming window to reduce spectral leakage.

3. Fourier Transform and Power Spectrum:

Each framed segment undergoes a Fast Fourier Transform (FFT) to obtain the magnitude spectrum.

The power spectrum is then calculated as the squared magnitude of the FFT coefficients.

4. Filter Banks (FB):

The Mel-scale filter banks are computed to simulate thehuman auditory system's response to different frequencies.

This involves mapping the linear frequency scale into the Mel scale, which is perceptually more meaningful.

Filter banks are applied to the power spectrum to extract relevant frequency components.

5. Mel-Frequency Cepstral Coefficients (MFCC):

MFCCs are derived from the logarithm of the filter bank outputs.

The Discrete Cosine Transform (DCT) is applied to the log filter bank energies to decorrelate the features.

The resulting coefficients (typically the first 13) represent a compact representation of the spectral envelope of the audio signal.

*Need for FB and MFCC in Language Detection:*

1. Discriminative Features:

FB and MFCC capture essential spectral information that characterizes speech sounds.

They are robust features that can differentiate between different phonemes and speech patterns across languages.

2. Dimensionality Reduction:

MFCC reduces the dimensionality of the audio signal while preserving relevant information.

This reduction simplifies the subsequent modelling and classification tasks.

3. Invariance to Noise and Speaker Variability:

FB and MFCC are relatively invariant to variations in noise and speaker characteristics.

They focus on the spectral content that is essential for phonetic distinctions across languages.

4. Modelling and Classification:

FB and MFCC serve as input features for machine learning models (e.g., neural networks, SVMs) used in language detection.

They facilitate accurate classification by capturing the underlying phonetic and linguistic characteristics of speech.

In summary, generating Filter Banks and Mel-Frequency Cepstral Coefficients from audio signals plays a crucial role in speech analysis and language detection tasks. These features enable effective representation of speech signals, aiding in the development of robust and accurate language detection systems.

*F. Model Development:*

The model development process involved training a Convolutional Neural Network with the aim of identifying the spoken language of an audio.

*Network Architecture:*

Input Layer:

The network takes in a 2D input in a single channel (grayscale).

Convolutional Layers:

2D convolution with 32 filters of size (7, 7), using ReLU activation and valid padding. This layer processes the input images to extract features.

BatchNormalization:

Normalizes the activations of the previous layer.

MaxPooling2D:

Performs max pooling with a pool size of (3, 3) and a stride of 2, reducing spatial dimensions while retaining important features.

Additional Convolutional Layers:

Similar structure with increasing filter sizes and deeper layers (64, 128, 256, 512) to capture hierarchical features.

Flatten Layer:

Flattens the output of the last convolutional layer into a 1D array, preparing it for input into a fully connected layer.

Fully Connected Layers:

A fully connected layer with 256 units and ReLU activation, followed by BatchNormalization and Dropout for regularization.

Final dense layer with 3 units (assuming 3 classes for classification) and softmax activation for multi-class classification output.

The developed CNN architecture is designed for a classification task. It consists of multiple convolutional layers followed by fully connected layers for feature extraction and classification. The model is trained using the Adam optimizer with dynamic learning rate adjustment and is evaluated based on its performance on test data. The use of BatchNormalization, Dropout, and callbacks like ModelCheckpoint and LearningRateScheduler enhances model performance, regularization, and optimization during training.

V. ANALYSIS OF RESULTS AND CONCLUSION

The model exhibits inadequate performance in recognizing Spanish language compared to its recognition of English and German.

The obtained confusion matrix reveals the model's accuracy and errors in distinguishing between English and Deutsch. The performance of the model on recognising English and Deutsch is given in the form of a confusion matrix obtained.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.48 | 0.30 | 0.37 | 122 |
| 1 | 0.40 | 0.93 | 0.56 | 122 |

Fig 5: Confusion matrix

The average accuracy achieved by the model is 43%, indicating a moderate level of performance. Interestingly, the model exhibits better proficiency in recognizing Deutsch compared to English, although there is still room for improvement in overall performance.

To enhance the model's capabilities, we can revisit and refine its architecture based on the latest advancements and insights in the field of speech recognition. By incorporating updated techniques and leveraging new information, adjustments can be made to the model's design to potentially enhance its accuracy and effectiveness across different languages, particularly in improving its performance for English recognition. This iterative process of refining the model's architecture ensures that it remains aligned with the cutting-edge developments in speech recognition technology, ultimately aiming for higher accuracy and robust language identification capabilities.

REFERENCES

[1] M. A. Shahriar, T. Aziz, S. Banik and A. Sattar, "Identification of Spoken Language using Machine Learning Approach," 2020 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392744.

[2] J. S. Anjana and S. S. Poorna, "Language Identification From Speech Features Using SVM and LDA," 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2018, pp. 1-4, doi: 10.1109/WiSPNET.2018.8538638.

[3] M. Ablimit, M. Xueli and A. Hamdulla, "Language Identification Research Based on Dual Attention Mechanism," 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 2021, pp. 241-246, doi: 10.1109/PRML52754.2021.9520699.

[4] H. Venkatesan, T. V. Venkatasubramanian and J. Sangeetha, "Automatic Language Identification using Machine learning Techniques," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 583-588, doi: 10.1109/CESYS.2018.8724070.

[5] A. R. Ambili and R. C. Roy, "Spoken Language Identification of Indian Languages in Adversarial Synthetic and Noisy Attacking Environments," 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), Kochi, India, 2022, pp. 1-6, doi: 10.1109/IC3SIS54991.2022.9885560.