Analyzing the language confusion pattern and mutual influences using deep neural network based spoken language identification systems

Internship completion report submitted in partial fulfillment of the requirements for the degree

of

Masters of Technology

by

Bhagyamuni Rajak
23EC65R05

Under the guidance of

Prof. Goutam Saha

ELECTRONICS & ELECTRICAL COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Submitted July 30, 2024

Declaration

I hereby declare that:

(a) The work presented in this thesis has not been submitted to any other Institute/University for any degree or diploma.

(b) I have adhered to the norms and guidelines outlined in the Ethical Code of Conduct of the Institute.

(c) Whenever materials such as data, theoretical analyses, figures, and text from other sources have been utilized, proper credit has been given to them by citing them in the text of the thesis and providing their details in the references. Additionally, I have obtained the necessary permissions from the copyright owners of these sources, as required.

Date: July 30, 2024 Bhagyamuni Rajak

Place: Kharagpur 23EC65R05 Project Guide Signature

Department of ELECTRONICS & ELECTRICAL COMMUNICATION
ENGINEERING
Indian Institute of Technology, Kharagpur
India - 721302

INTERNSHIP COMPLETION CERTIFICATE

Date : 30/07/2024
Place : Kharagpur

This is to certify that **Bhagyamuni Rajak**(Roll Number: 23EC65R05) has completed an internship under my guidance and supervision during the period 15/05/2024 to 30/07/2024 on ”**Analyzing the language confusion pattern and mutual influences using deep neural network based spoken language identification systems**”, as a part of her master's degree program. The undersigned may be contacted for further information about the internship.

Signature : ....
Prof. Goutam Saha

ACKNOWLEDGEMENT

I sincerely thank Prof. Goutam Saha for their constant guidance, support, and constructive feedback throughout the project. His advice and concepts helped carry out the thesis work correctly. His unwavering encouragement and inspiration helped me push through all the hurdles and get to this stage.

I am very much thankful to Prof. Tarun Kanti Bhattacharyya, head of the E&ECE department at IIT Kharagpur, for providing the required facilities during the thesis study

Last, but not least, I also want to thank my friends and colleagues, my parents, and respected seniors for their support and helping hands whenever I needed them.

Bhagyamuni Rajak
M.Tech (VIS)
Indian Institute of Technology, Kharagpur
Date:30/07/2024

ABSTRACT

A language recognition system using the ECAPA-TDNN neural network architecture. Leveraging the IIITH-ILSC dataset, which includes 23 Indian languages, the system employs advanced feature extraction methods such as Mel Frequency Cepstral Coefficients (MFCCs) and wav2vec 2.0. The results demonstrate that the ECAPA-TDNN model, especially with wav2vec 2.0 features, significantly improves language recognition accuracy compared to traditional methods like GMM-UBM and SVM. This underscores the effectiveness of neural networks in language identification and provides insights into cross-linguistic influences among Indian languages.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Language Identification (LID) systems aim to recognize the language of spoken utterances. The development of spoken dialog systems has garnered significant interest, especially for Indian languages due to their similar origins and overlapping phonesets[5]. This makes developing LID systems for these languages particularly challenging. Despite these similarities, Indian languages differ significantly in phonotactics, prosody, and intonations. The need for LID systems arises from their applications in various speech technologies such as multilingual dialog systems and information-query systems. In a multilingual automatic speech recognition (MASR) system, LID acts as a front-end to switch between multiple monolingual ASR systems[1]. When used as a front-end switch in dialog systems, LID can help the systems operate more robustly by leveraging phonotactic constraints. Voice-operated applications, such as automatic call routing, also rely on LID to route incoming calls to the appropriate application.

1.2 Language Families

Language families are groups of languages that share a common ancestral language. They are categorized based on linguistic features such as phonetics, syntax, morphology, and vocabulary[5]. Understanding language families helps linguists trace the evolution of languages, uncover historical connections among cultures, and provide insights into the migration patterns of ancient populations.

Major Language Families

Globally, there are several major language families, each with numerous languages under its umbrella. Two prominent language families in the Indian subcontinent are the Indo-European and Dravidian families as shown in (Fig. 1.1).

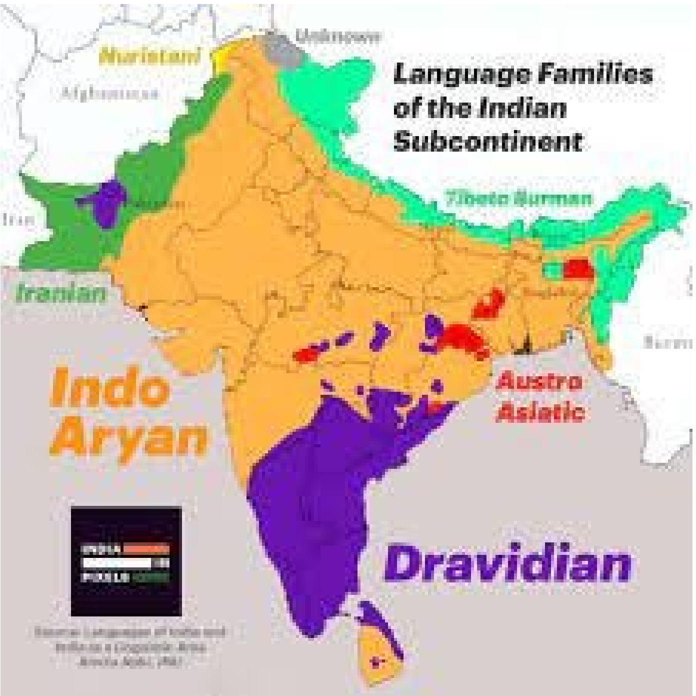


Figure 1.1: Indian Language Families.

The Indo-European language family is one of the largest and most widely dispersed language families in the world. It includes languages spoken in Europe, parts of Asia, and the Americas. Within India, the Indo-European family encompasses many languages predominantly spoken in Central and Northern India.
Notable Indo-European Languages in India: Hindi, Bengali, Marathi, Gujarati, Punjabi, Odia, and Konkani. Historical Influence: These languages have evolved from ancient languages like Sanskrit and have been influenced by Persian, Arabic, and other regional languages over time. The Dravidian language family is primarily found in Southern India and parts of Eastern and Central India. It comprises languages that are rich in literature and have ancient roots predating the Indo-European influence in the region. Dravidian Languages: Tamil, Telugu, Kannada, and Malayalam. Cultural Significance: These languages have their own scripts and have contributed significantly to Indian literature, art, and culture.

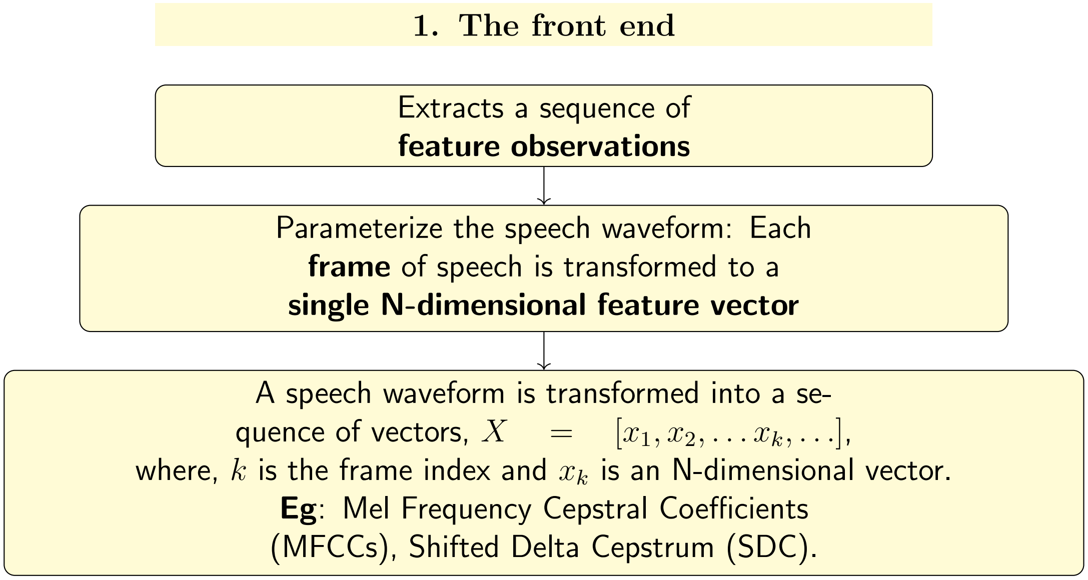
Cross-Linguistic Influence

The interaction between Indo-European and Dravidian languages has led to significant cross-linguistic influence, resulting in linguistic features being shared and adapted among these language families.Indo-European Influence on Dravidian Languages: Indo-European languages have influenced Dravidian languages in terms of vocabulary, phonetics, and syntax, especially in regions where these languages coexist. Dravidian Influence on Indo-European Languages: Dravidian languages have similarly impacted Indo-European languages, contributing unique phonetic and syntactic features.

1.3 Language Identification Systems

Language Identification (LID) systems are designed to automatically recognize the language of spoken utterances. These systems are essential in various applications such as multilingual dialog systems, automatic speech recognition (ASR), and voice-operated services like call routing.
The LID system is divided into two main sections as shown in (Fig. 1.2) and (Fig. 1.3) Front-End: Extracts a sequence of feature observations, transforming each frame of speech into an N-dimensional feature vector. Common methods include Mel Frequency Cepstral Coefficients (MFCCs) and Shifted Delta Cepstrum (SDC). Back-End: Involves model training using feature vectors to train a separate model for each language. During the identification phase, similar speech features are extracted from unknown utterances and compared to the trained models[1].

The LID system is divided into two sections :



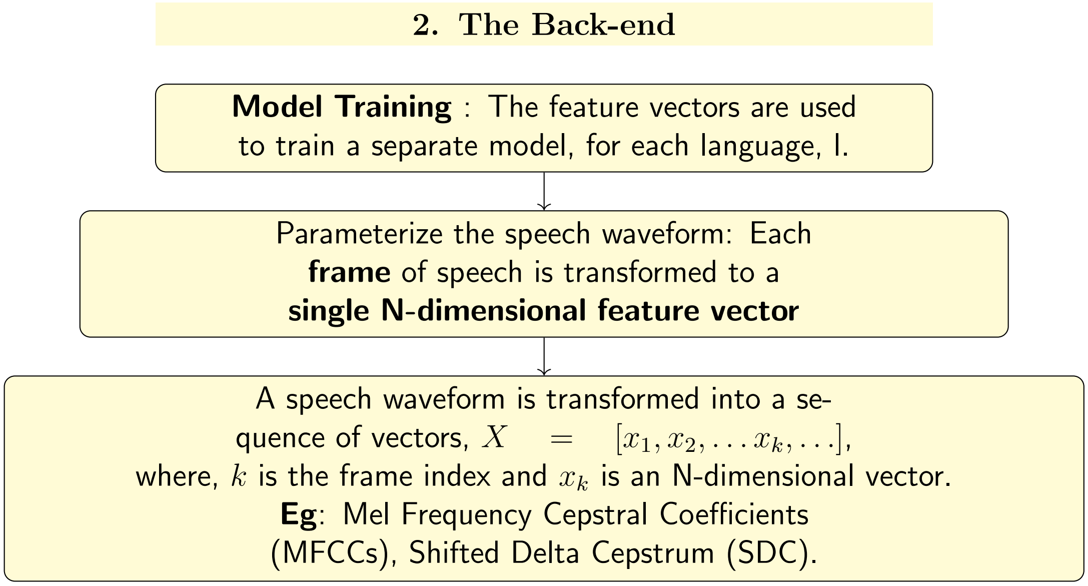
1. The front end

Extracts a sequence of **feature observations**

Parameterize the speech waveform: Each **frame** of speech is transformed to a **single N-dimensional feature vector**

A speech waveform is transformed into a sequence of vectors, $X = [x_1, x_2, \ldots x_k, \ldots]$, where, k is the frame index and x_k is an N-dimensional vector. **Eg**: Mel Frequency Cepstral Coefficients (MFCCs), Shifted Delta Cepstrum (SDC).

Figure 1.2: Front-end of the LID system



2. The Back-end

Model Training : The feature vectors are used to train a separate model, for each language, l.

Parameterize the speech waveform: Each **frame** of speech is transformed to a **single N-dimensional feature vector**

A speech waveform is transformed into a sequence of vectors, $X = [x_1, x_2, \ldots x_k, \ldots]$, where, k is the frame index and x_k is an N-dimensional vector. **Eg**: Mel Frequency Cepstral Coefficients (MFCCs), Shifted Delta Cepstrum (SDC).

Figure 1.3: Back-end of the LID system

1.4 Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network.

The **ECAPA-TDNN** (Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network) model[2] enhances traditional TDNNs with advanced channel attention, propagation, and aggregation mechanisms for superior speaker and language recognition. Emphasized channel attention, using Squeeze-Excitation (SE) blocks, refines feature focus for better discrimination. Propagation and aggregation processes capture long-term dependencies and ensure comprehensive temporal context. Overall, ECAPA-TDNN offers significant improvements in feature extraction and model performance. Below is a detailed description of the ECAPA-TDNN architecture, accompanied by (Fig. 1.4)
The architecture includes 1-dimensional SE-Res2Blocks, which integrate global properties into frame-level features by averaging inputs (”squeeze”) and applying learned weights (”excite”). Res2Net modules within SE-Res2Blocks process multi-scale features, enhancing pattern capture and stability while reducing overfitting. Multi-layer feature aggregation combines shallow and deep features from all SE-Res2Blocks for robust speaker embeddings. Each frame layer block uses the output of all preceding blocks as input, ensuring comprehensive feature integration[4]. Features are combined through concatenation and summation to balance parameter count and enhance capacity. Attentive statistics pooling captures frame-level variations and speaker traits using channel-wise normalization and weighted mean and standard deviation. Aggregated features are processed through a dense layer with Additive Angular Margin (AAM) Softmax for classification, improving discriminative power. Key parameters

include kernel size, dilation, channel dimension, temporal dimension, and the number of training speakers, which all impact the model's performance.robust embeddings.

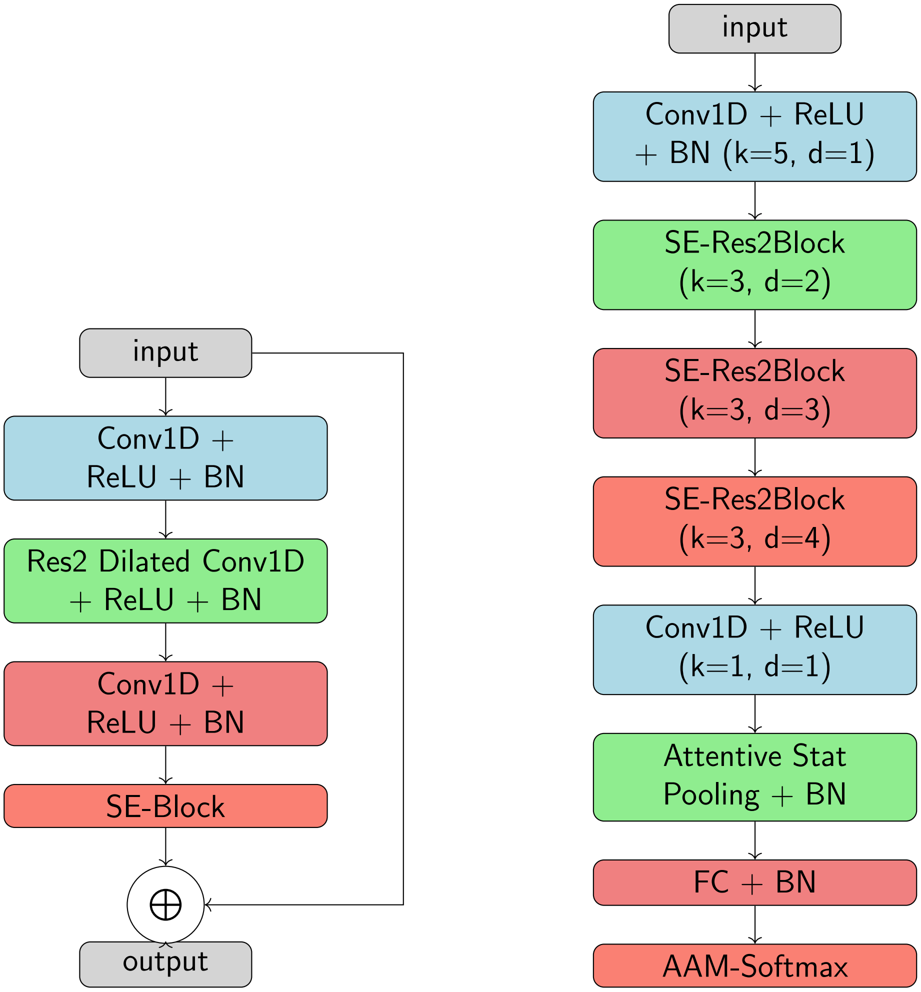


Figure 1.4: Network topology of ECAPA-TDNN

Chapter 2

Literature Survey

Language Identification (LID) has significantly evolved, tackling the challenge of recognizing spoken languages through feature extraction and model construction[5]. Traditional methods like Gaussian Mixture Models (GMM), Universal Background Models (UBM), and Support Vector Machines (SVM) laid the groundwork but struggle with the complexity and variability of natural languages, especially in India's diverse linguistic context[6]. Recent advancements in neural network architectures, particularly the ECAPA-TDNN (Emphasized Channel Attention Propagation and Aggregation Time Delay Neural Network) model, have significantly improved language recognition systems[2]. ECAPA-TDNN enhances feature extraction and recognition performance through advanced attention and aggregation mechanisms. Integrating deep learning techniques into LID systems has deepened understanding of cross-linguistic influences and improved model development[4], with studies examining interactions between Dravidian and Indo-European languages to improve recognition accuracy[5]. Current research focuses on developing and fine-tuning models like ECAPA-TDNN for effective language recognition across diverse language families, using balanced datasets from both families to ensure comprehensive training[6]. Comparative analyses show significant improvements over traditional methods. The goal is to enhance language recognition systems for Indian languages, providing better cross-linguistic understanding and advancing speech recognition technologies for multilingual applications[3].

Chapter 3

Problem Statement

Language is a structured system of communication, and language families are groups of related languages. India, the most linguistically diverse nation, hosts languages primarily from the Indo-European and Dravidian families. These languages' mutual influence offers insights into historical and socio-political contexts. This study aims to leverage our model to assess linguistic closeness among Indian languages by training Language Identification (LID) models. By examining model performance with and without certain languages, we can explore linguistic relationships and influences. This innovative approach, limited in current LID literature and rarely conducted using deep learning, will enhance understanding of cross-linguistic influences and demonstrate our model's potential in linguistic research.

We aim to develop different state-of-the-art LID frameworks and study their confusion matrices to find out which Indian languages are most confused with each other. To ensure the linguistic observation, we alter the features, databases, features and study the language relations by confusion matrices. With this as an motivation, we then further extend this study to Indian language family identification and finding out the influence of one language family to another considering both neighboring and non-neighbourin languages.

Chapter 4

Solution

To address the problem of understanding linguistic diversity and influences among Indian languages, we leveraged advanced neural network architectures, feature extraction techniques, diverse datasets, and analysis of language families to achieve high performance in language identification. Our research integrated and fine-tuned the ECAPA-TDNN model for Indian languages, adjusting hyperparameters and architecture to enhance accuracy. Combining traditional MFCC and modern wav2vec2.0 feature extraction methods enabled the model to discern subtle linguistic nuances, while diverse datasets (IIITH, LDC, KGP) ensured better generalization. The model's robustness was demonstrated by its ability to identify language families and handle mixed-language scenarios, even with languages not included in training. Additionally, our detailed analysis of cross-linguistic dynamics between Indo-European and Dravidian languages provided insights into the complexities of language interactions and their impact on recognition systems. This study not only enhances understanding of linguistic relations but also sets new benchmarks in speech processing, showcasing the potential of deep learning to expand traditional linguistic research.

Chapter 5

Implementation

5.1 Dataset Preparation

5.1.1 Indian Language Speech Corpus

Details about the speech corpus are described in (Table 5.1).After collecting the data each speech file is chopped into 5-10 sec utterances using Audacity software. While chopping the record, the utterance with background music, non-speech sounds are avoided in the preparation of database. Most of the data which is collected from the internet and broadcasts are verified by the native speaker of corresponding languages. The speech corpus is pooled from various sources with files of different formats and sampling rates. For the uniformity, all files are converted into .wav format with a sampling rate of 8000 samples/sec using Audacity software.
Next, we calculated the Mel Frequency Cepstral Coefficients (MFCCs), a widely used filter-bank parameterization method. MFCCs approximate the nonlinear frequency resolution of the human ear using the Mel scale. The calculation involves computing the magnitude-square of the Fourier Transform for the input windowed speech frame, passing the result through triangular Mel filters, and taking the natural logarithm of the filter bank energies. To decorrelate the highly correlated filter bank log-energies, a Discrete Cosine Transform (DCT) is applied.
The process starts by reading the speech file from the specified file path, sampling it at the given fs, and dividing the signal into frames based on Window Length. The Fourier Transform is applied with NFFT bins, and the result is passed through No Filter Mel filter banks. Energy based VAD with the threshold th is used to detect voiced frames. The final output, t, consists of the static MFCC features for these frames.

Table 5.1: IIITH-ILSC Dataset Overview

IIITH-ILSC Dataset Overview	
Dataset	IIITH-ILSC (Indian Language Speech Corpus)
Languages	23 Indian languages including 22 official languages and English.
Languages Covered	Assamese, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Urdu, Indian English.
Sources	Archives of Prasar Bharati, All India Radio, TED-talks, broadcast speech, and university students.
Total Speakers	1150 (50 per language, 25 male, 25 female)
Total Hours	103.5 hours
Training & Testing	3.5 hours for training, 1 hour for testing per language.
Speech Quality	60% clean, 40% moderately noisy.
Format	.wav at 8000 samples/sec.

5.1.2 Cross-Corpus Dataset

Cross-corpora linguistic relation refers to the study of how well language identification (LID) systems perform when they are trained on one dataset (corpus) and tested on another[3]. This is important because real-world scenarios often involve diverse audio sources, and a robust LID system must be able to handle this diversity effectively.

Table 5.2: Comparison of Speech Corpora

Corpora	**IIITH**	**LDC**	**KGP**
Total languages	23	5	27
Mode of speech	BN and CTS	CTS	BN
Environment	Studio, real-world	Real-world	Studio
Total speakers	1150	584	300
Duration	103.5 hours	118.3 hours	27 hours
Audio format	16 kHz (.wav)	8 kHz (.flac)	8 kHz (.wav)

5.2 ECAPA-TDNN Model for 23 Language Recognition

The ECAPA-TDNN model has been implemented for the challenging task of recognizing 23 different languages, the ECAPA-TDNN model's architecture is designed to capture temporal dependencies and enhance feature discrimination for robust language recognition.
Architecture Overview:
Input Layer: The model starts with an input layer that processes language features over time. The input is a feature matrix of size 80×T, where 80 is the number of Mel-frequency cepstral coefficients (MFCCs) or other feature dimensions, and T is the temporal dimension (time frames).
Conv1D + ReLU + BN Layer: A 1D convolutional layer followed by ReLU activation and Batch Normalization (BN) to extract initial features. Conv1D: This performs a 1D convolution operation with kernel size k=5 and dilation d=1. The convolution operation can be represented as:

$$y[t] = \sum_{i=0}^{k-1} x[t + i \cdot d] \cdot w[i] \tag{5.1}$$

where x is the input, y is the output, w is the filter, and t is the time index.
ReLU (Rectified Linear Unit): Applies the activation function:

$$ReLU(x) = max(0, x) \tag{5.2}$$

BN (Batch Normalization): Normalizes the activations to improve training stability and performance.
SE-Res2Blocks: Three Squeeze-and-Excitation Res2Blocks with varying kernel sizes and dilation spacings to capture multi-scale temporal features.Performs channel-wise feature recalibration. It consists of: Squeeze: Global average pooling to generate channel-wise statistics:

$$z_c = \frac{1}{T} \sum_{t=1}^{T} x_{c,t} \tag{5.3}$$

Excitation: A two-layer fully connected network to model channel dependencies:

$$s_c = \sigma(W_2 \cdot ReLU(W_1 \cdot z)) \tag{5.4}$$

where W_1 and W_2 are weights, and σ is the sigmoid function. Each SE-Res2Block

includes Conv1D layers, dilation, ReLU activation, and SE-Block to enhance feature recalibration.The SE-Res2Block in the (Fig. **??**) shows different configurations of kernel sizes (k) and dilation rates (d) to capture various temporal dependencies.
Attentive Statistics Pooling Layer: Aggregates temporal features into a fixed-size representation, incorporating attention mechanisms to focus on important features, and aggregates the frame-level features into a single vector by weighted averaging, where the weights are learned through an attention mechanism:

$$attention(h_t) = \frac{exp(W_a h_t)}{\sum_{t=1}^{T} \exp(w_a h_t)} \tag{5.5}$$

The weighted average is:

$$v = \sum_{t=1}^{T} attention(h_t) \cdot h_t \tag{5.6}$$

Fully Connected (FC) + BN Layer: A fully connected layer followed by BN to project pooled features into the final embedding space.Fully Connected (FC) Layer maps the pooled feature vector to a fixed-size output. BN (Batch Normalization) normalizes the output of the FC layer.
AAM-Softmax Layer: Additive Angular Margin Softmax (AAM-Softmax) for classification, introducing an angular margin to improve class separation. A variant of softmax used for improving the discriminative power of the embeddings by adding an angular margin:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s \cdot (\cos(\theta_{y_i} + m)))}{\exp(s \cdot (\cos(\theta_{y_i} + m))) + \sum_{j \neq y_i} \exp(s \cdot \cos(\theta_j))} \tag{5.7}$$

where N is the batch size, s is a scaling factor that controls the magnitude of the logits, m is the angular margin that enforces extra margin between classes, θ_{y_i} is the angle between the feature vector of the i-th sample and the weight vector of the correct class, θ_j is the angle between the feature vector of the i-th sample and the weight vector of the j-th class, and y_i is the true label of the i-th sample.

During training, language data is processed through the model to compute loss and update weights, minimizing errors. Data is divided into mini-batches for efficient processing: larger batches leverage parallel processing and hardware acceleration, while smaller batches offer frequent updates at higher computational cost. The validation phase assesses model performance on unseen data to ensure generalization, using a separate validation set without updating weights. Performance is measured through metrics like accuracy, precision, recall, F1-score, EER, and validation loss.

5.3 Wav2Vec 2.0 feature extractor

Self-supervised learning involves training models on a large amount of unlabeled data by creating auxiliary tasks that generate labels from the data itself. In the context of Wav2Vec 2.0, the model learns to predict masked parts of the speech signal from unmasked parts, similar to the masked language modeling used in BERT for text. The input to the model is the raw audio waveform, denoted as X. The raw waveform X is first passed through a series of convolutional layers to extract latent speech representations Z. These layers act as a feature extractor that captures local patterns in the waveform. Mathematically: Let f_{CNN} represent the convolutional layers, then:

$$Z = f_{\text{CNN}}(X)$$

The latent speech representations Z are then quantized to obtain quantized representations Q. This step involves a vector quantization (VQ) process where the continuous latent vectors are mapped to discrete codes from a learned codebook. Mathematically: Let $\mathcal{Q}$ be the quantization function, then:

$$Q = \mathcal{Q}(Z)$$

The model then uses a Transformer network to produce context representations C from the latent speech representations Z. The Transformer captures long-range dependencies and contextual information from the speech signal. Mathematically: Let $f_{\text{Transformer}}$ represent the Transformer network, then:

$$C = f_{\text{Transformer}}(Z)$$

A contrastive loss is applied to ensure that the context representations C are informative about the quantized representations Q. The goal is to maximize the similarity between C and the corresponding Q while minimizing the similarity between C and negative samples. Mathematically: The contrastive loss $\mathcal{L}_{\text{contrastive}}$ can be expressed as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(C, Q))}{\sum_{Q' \in \text{Negatives}} \exp(\text{sim}(C, Q'))} \tag{5.8}$$

where sim is a similarity function (e.g., cosine similarity) and Q' are negative samples. Finally, the context representations C are linearly projected to the output Y. Mathematically: Let W be the weight matrix for the linear projection, then:

$$Y = WC$$

Chapter 6

Results

We investigate the linguistic relation in two ways:

- First, we train a LID and use its evaluation set confusion matrix to assess how one language is confused with another.
  The confusion matrix in the **Fig. 6.1** shows the performance of the model across 23 different languages. Diagonal elements of the confusion matrix represent correct predictions, with higher values indicating strong performance for that language. Off-diagonal elements indicate misclassifications. The model's overall accuracy is 71.29%. Hindi shows high accuracy due to its distinct features and widespread use, while Gujarati and Marathi have noticeable misclassifications. Marathi, being closer to Dravidian languages, faces more misclassifications due to linguistic similarities.

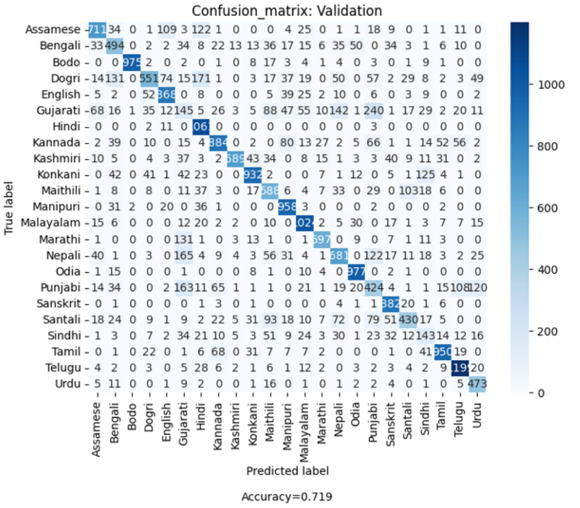
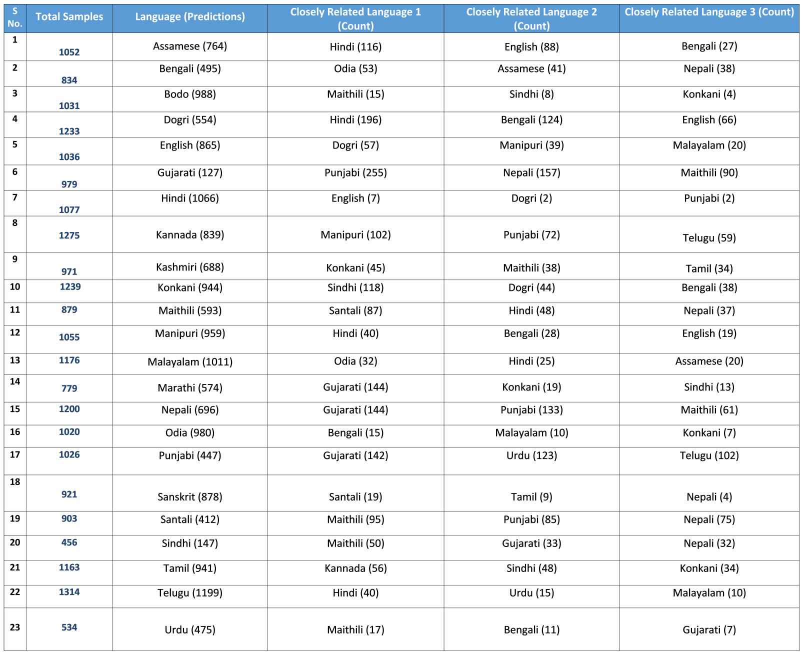


Figure 6.1: Overview of the ECAPA-TDNN model implementation for 23 language recognition.



S No.	Total Samples	Language (Predictions)	Closely Related Language 1 (Count)	Closely Related Language 2 (Count)	Closely Related Language 3 (Count)
1	1052	Assamese (764)	Hindi (116)	English (88)	Bengali (27)
2	834	Bengali (495)	Odia (53)	Assamese (41)	Nepali (38)
3	1031	Bodo (988)	Maithili (15)	Sindhi (8)	Konkani (4)
4	1233	Dogri (554)	Hindi (196)	Bengali (124)	English (66)
5	1036	English (865)	Dogri (57)	Manipuri (39)	Malayalam (20)
6	979	Gujarati (127)	Punjabi (255)	Nepali (157)	Maithili (90)
7	1077	Hindi (1066)	English (7)	Dogri (2)	Punjabi (2)
8	1275	Kannada (839)	Manipuri (102)	Punjabi (72)	Telugu (59)
9	971	Kashmiri (688)	Konkani (45)	Maithili (38)	Tamil (34)
10	1239	Konkani (944)	Sindhi (118)	Dogri (44)	Bengali (38)
11	879	Maithili (593)	Santali (87)	Hindi (48)	Nepali (37)
12	1055	Manipuri (959)	Hindi (40)	Bengali (28)	English (19)
13	1176	Malayalam (1011)	Odia (32)	Hindi (25)	Assamese (20)
14	779	Marathi (574)	Gujarati (144)	Konkani (19)	Sindhi (13)
15	1200	Nepali (696)	Gujarati (144)	Punjabi (133)	Maithili (61)
16	1020	Odia (980)	Bengali (15)	Malayalam (10)	Konkani (7)
17	1026	Punjabi (447)	Gujarati (142)	Urdu (123)	Telugu (102)
18	921	Sanskrit (878)	Santali (19)	Tamil (9)	Nepali (4)
19	903	Santali (412)	Maithili (95)	Punjabi (85)	Nepali (75)
20	456	Sindhi (147)	Maithili (50)	Gujarati (33)	Nepali (32)
21	1163	Tamil (941)	Kannada (56)	Sindhi (48)	Konkani (34)
22	1314	Telugu (1199)	Hindi (40)	Urdu (15)	Malayalam (10)
23	534	Urdu (475)	Maithili (17)	Bengali (11)	Gujarati (7)

Figure 6.2: Analysis of Language Classification Model Confusion Matrix.

- For each language, we identified the top-3 confusing languages. The **Fig. 6.2** shows the confusion matrix of a language classification model, detailing the total samples, predicted languages, and the top three frequently mispredicted languages. Hindi, with 1077 samples, is the most predicted language. This confusion arises from linguistic similarities within language families or geographic regions, shared scripts and orthography. These similarities within language families, shared scripts, code-switching, and potential mislabeling in training data contribute to the misclassifications. By addressing these factors we can improve model accuracy.

  To verify if such relations are based on language or just dependent on LID system configuration we did several experiments.

- We train with different Features: The model trained with wav2vec2.0 features shows significantly higher accuracy than the MFCC-based model, capturing more relevant features for language recognition. The confusion matrix in (Fig. 6.3) shows high correct prediction rates, indicating strong language identification capabilities with an overall accuracy of 93.1%. Fewer significant off-diagonal elements highlight challenges in distinguishing closely related languages. Thus, the ECAPA-TDNN model with wav2vec2.0 features achieves exceptional accuracy and robustness in language recognition tasks.

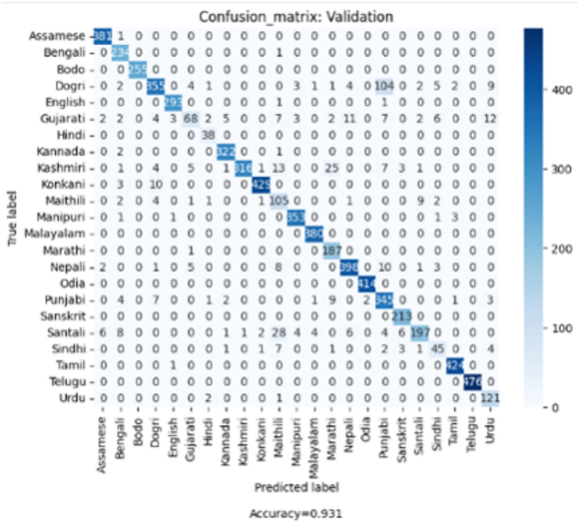
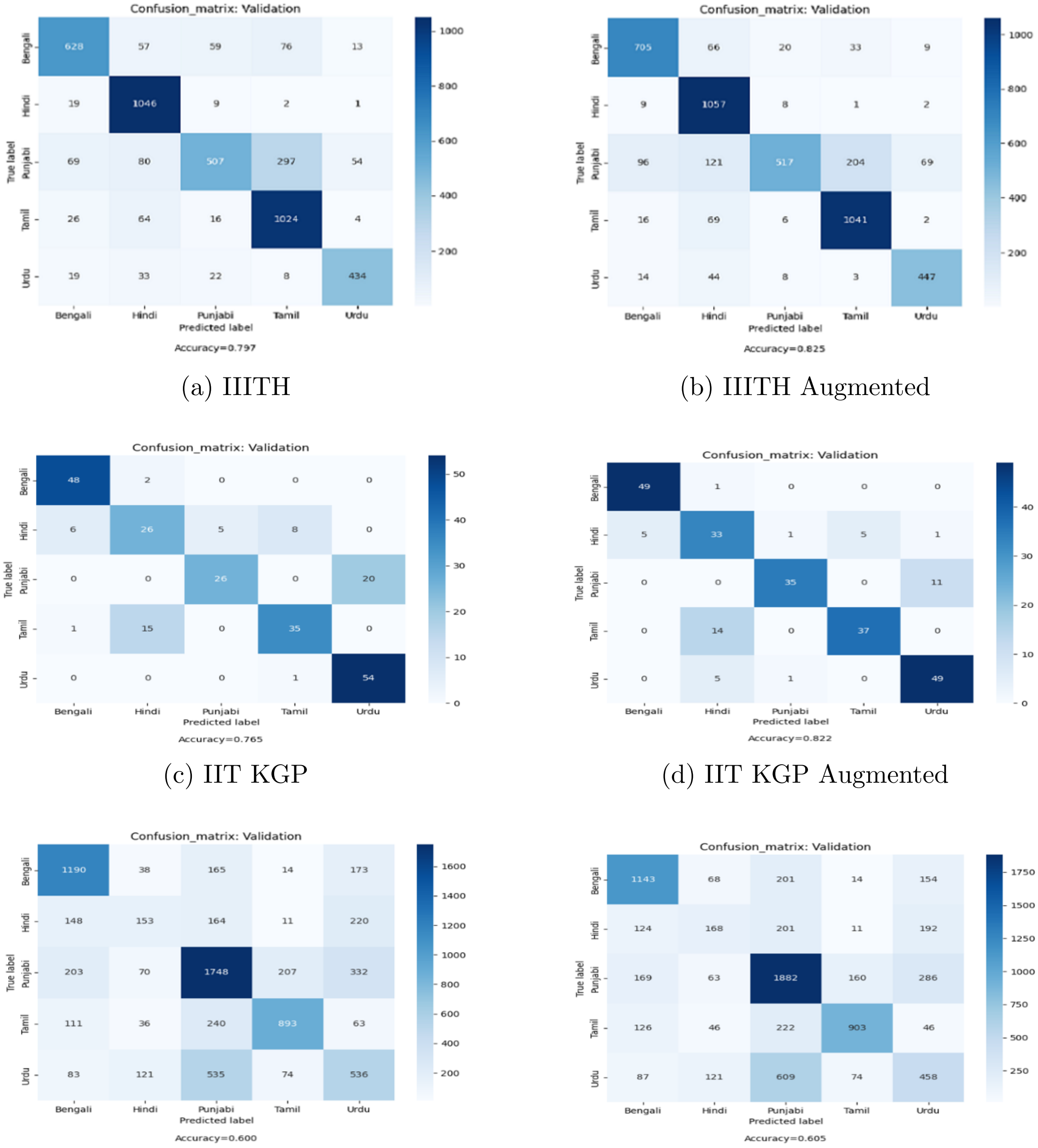


Figure 6.3: Overview of the ECAPA-TDNN model implementation for 23 language recognition by wav2vec2.0 feature.

- We train with different Databases: We evaluated the performance of Language Identification (LID) systems trained on three distinct corpora: IIITH-ILSC (IIITH), LDC South Asian (LDC), and IITKGP-MLILSC (KGP). Testing within and across these corpora revealed that the LDC-tested model had lower validation accuracy as shown in **Fig. 6.4** due to significant non-lingual variations, highlighting challenges in real-world generalization. The IIITH-trained model outperformed on the KGP corpus, indicating similarities between IIITH and KGP data. To address corpora-mismatch, environmental augmentations were used during training, simulating different recording conditions, which successfully reduced mismatch and improved the model's generalization ability.

  The IIITH corpus yields high accuracy, indicating effective learning and adaptation by the model.The LDC corpus presents significant challenges due to its variations, resulting in lower accuracy and demonstrating the model's struggle to generalize to vastly different data.
  The KGP corpus shows improved accuracy with augmentation, emphasizing the benefit of exposing the model to diverse training scenarios to enhance generalization.These results highlight the importance of training on diverse datasets and employing environmental augmentations to develop robust and adaptable language identification systems

- The consistent misprediction patterns across different corpora reveal intriguing relationships between languages. For instance, Bengali often gets misclassified as Hindi, suggesting inherent similarities between the two. This pattern holds



(a) IIITH

(b) IIITH Augmented

(c) IIT KGP

(d) IIT KGP Augmented

(e) LDC

(f) LDC Augmented

Figure 6.4: The Cross-Corpora test results

Table 6.1: Accuracy of Model Trained with IIITH Corpus and Tested with Different Corpora

Test Corpus	With Augmentation (%)	Without Augmentation (%)
IIITH	82.5	79.7
LDC	60.5	60.0
IITKGP	82.2	76.5

across various datasets, indicating that these languages share common characteristics. Notably, our model performs impressively well even with real-world datasets, reinforcing its robustness and the meaningful connections it identifies among languages.

- The second way of investigating linguistic similarity is based on finding impact of one language family on others.
  We consider only Indo-Aryan and Dravidian language families since they compose more than 90% speakers in India.We select top-4 most widely spoken Indo-Aryan and Dravidian languages. Among them two neighboring and two non-neighboring.

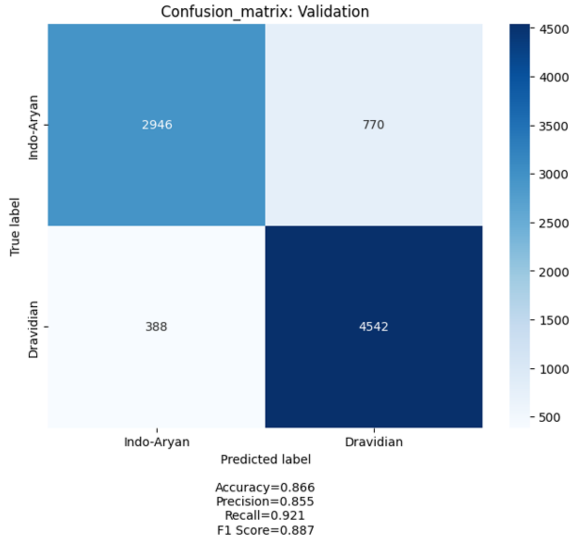
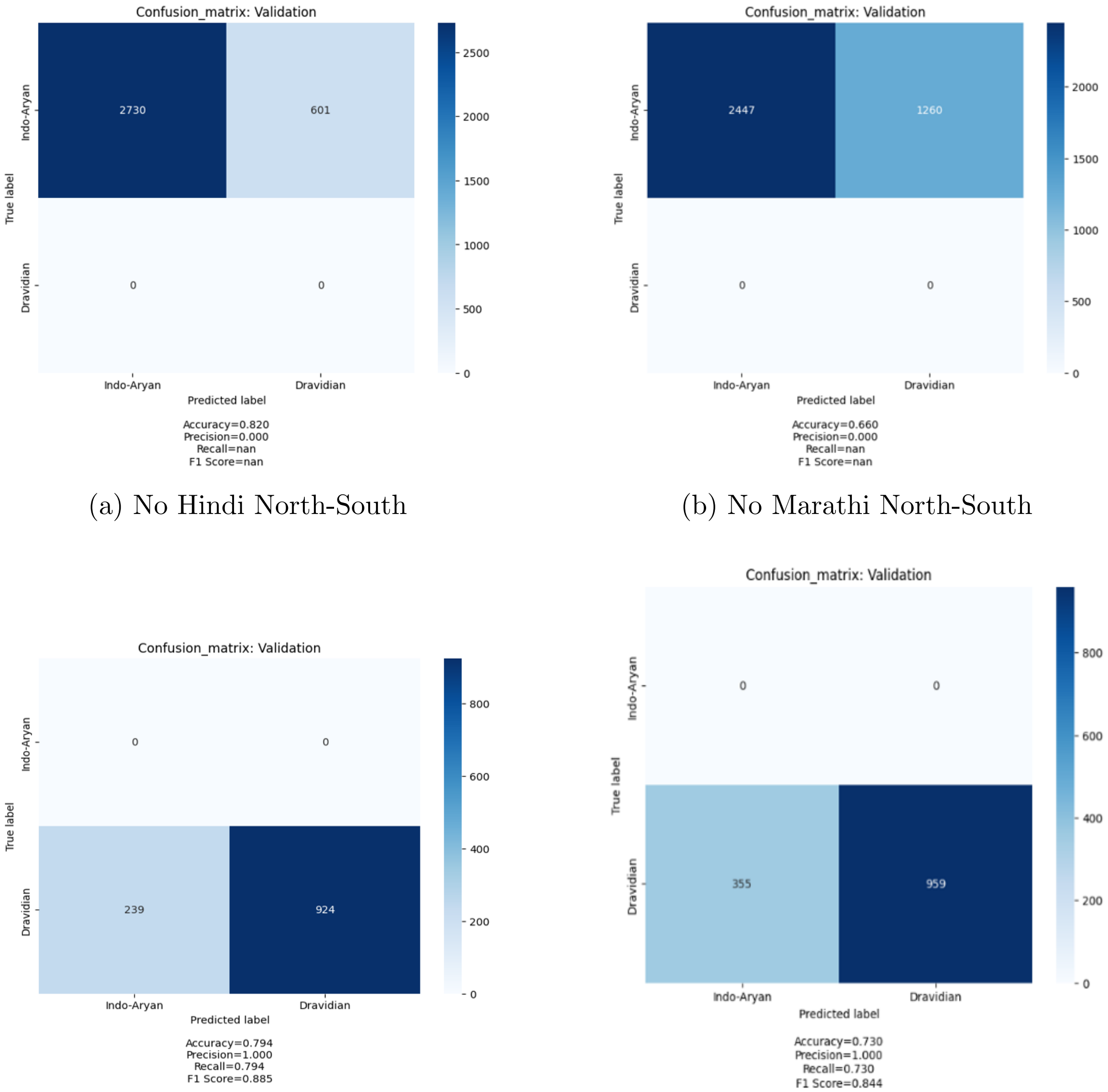


Figure 6.5: Language family identification.

- We proposed methodology how to assess the influence of one language family on the other.The (Fig. 6.6) shows the cross-linguistic influence between Indo-European and Dravidian languages is evident in model performance variations. By excluding Hindi and including another Indo-Aryan language while keeping



(a) No Hindi North-South

(b) No Marathi North-South

(c) No Tamil North-South

(d) No Telugu North-South

Figure 6.6: cross-linguistic influence between Indo-European and Dravidian languages.

Dravidian data constant, we assess the impact on accuracy. In our analysis, we focused on the most widely spoken languages from the Indo-European and Dravidian families, including Bengali, Punjabi, Kannada, and Malayalam. The validation accuracy for these languages is summarized in the (Table 6.2). Interestingly, languages that are geographical or linguistic neighbors to Dravidian or Indo-European languages, such as Marathi and Telugu, exhibit lower accuracy compared to non-neighboring languages.
This discrepancy in accuracy can be attributed to several factors, including historical interactions, geographical proximity, and inherent linguistic similarities.

Table 6.2: Influence of Language Families

INFLUENCE OF DRAVIDIAN FAMILY ON INDO-EUROPEAN LANGUAGES			
Hindi	Marathi	Bengali	Punjabi
Validation Accuracy: 82%	Validation Accuracy: 66%	Validation Accuracy: 79.1%	Validation Accuracy: 79.3%
INFLUENCE OF INDO-EUROPEAN FAMILY ON DRAVIDIAN LANGUAGES			
Tamil	Malayalam	Telugu	Kannada
Validation Accuracy: 79.4%	Validation Accuracy: 74.8%	Validation Accuracy: 73%	Validation Accuracy: 78.6%

Chapter 7

Conclusion

In our study, we delved into the fascinating world of linguistic diversity in India, one of the most linguistically rich nations globally, to understand how Indian languages influence one another. By leveraging Deep Learning-based Spoken Language Identification alongside traditional linguistic analysis, we aimed to uncover the relationships between languages from the Indo-Aryan and Dravidian families, which together account for over 90% of India's population. We trained Language Identification (LID) models and examined their confusion matrices to identify how one language is often mistaken for another, conducting various experiments to ensure the reliability of our findings. This involved analyzing different databases, features, and classifiers, revealing the top three languages each was most confused with. Additionally, we proposed a methodology to assess the impact of one language family on another, focusing on the most widely spoken languages from both families and distinguishing between neighboring and non-neighboring languages. Our findings indicate that neighboring languages exhibit more mutual influence, a conclusion supported by multiple feature analyses. This study not only enhances our understanding of linguistic relations but also offers insights into historical, socio-political contexts, showcasing the potential of Deep Learning to complement and expand traditional linguistic research.

References

[1] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2):82–108, 2011.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[3] Spandan Dey, Goutam Saha, and Md Sahidullah. Cross-corpora language recognition: A preliminary investigation with indian languages. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 546–550. IEEE, 2021.

[4] Lujun Li, Yikai Kang, Yuchen Shi, Ludwig Kürzinger, Tobias Watzel, and Gerhard Rigoll. Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021:1–16, 2021.

[5] Debapriya Sengupta and Goutam Saha. Identification of the major language families of india and evaluation of their mutual influence. *Current Science*, pages 667–681, 2016.

[6] Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, and Anil Kumar Vuppala. Iiith-ilsc speech database for indain language identification. In *SLTU*, pages 56–60, 2018.