Telugu Dialect Recognition Using CNN-LSTM Hybrid Architecture

Jaidev K, Jampala Sai Chandana, Munnangi Pranish Kumar, T Charishma Chowdary
Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India

***Abstract*—Telugu, an Indian language, is spoken by approximately 96 million people worldwide, primarily in the states of Andhra Pradesh and Telangana in India. Within these regions, Telugu exhibits four distinct dialects: Telangana (northern), Rayalaseema (southern), Coastal Andhra (central), and North Andhra (eastern). Understanding and classifying these dialects is crucial for natural language processing (NLP) and speech processing tasks. However, the presence of code-mixed data, where Telugu is mixed with other languages, complicates classification. To address this challenge, we collected code-mixed datasets from social media platforms, containing 2 hours of data for each dialect. We extracted MFCC (Mel Frequency Cepstral Coefficients) features from this data to capture speech characteristics. We employed a CNN-LSTM framework to leverage spatial and temporal features of speech signals for accurate classification. This framework integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to effectively process and classify code-mixed Telugu speech signals.**

***Index Terms*—Telugu, dialect recognition, code-mixed data, CNN-LSTM hybrid architecture, MFCC.**

I. Introduction

In the realm of burgeoning digital communication and the increasing integration of technology into our daily lives, the proliferation of speech data is experiencing an unprecedented surge. With the growing accessibility to digital platforms and communication channels, a diverse array of languages and dialects permeate the landscape of spoken interactions. Consequently, the accurate identification of languages assumes paramount importance in the domain of speech processing. Language detection, the fundamental task of determining the language spoken within a given audio segment, presents significant challenges, particularly in the context of multilingual speech environments.

Identifying comparable languages or dialects remains one of the most difficult tasks in language detection. This is most noticeable in Andhra Pradesh and Telangana in India, where Telugu, a Dravidian language, is spoken. Telugu has a vast linguistic variety, distinguished by unique regional accents and dialects. Before its breakup in 2014, Andhra Pradesh had four major regional dialects: Telangana in the north, Rayalaseema in the south, Coastal Andhra in the centre areas, and North Andhra in the east. Each dialect has distinct phonetic, lexical, and syntactic traits that reflect the region's cultural and geographical variety.

In addition to regional variations, Telugu often undergoes code-mixing with other languages like English, Hindi, or Urdu, particularly on social media platforms. This presents added hurdles for language identification algorithms, as they must navigate the complexities of multiple languages within a single document, compounded by the scarcity of available data. Developing a model capable of accurately classifying various dialects even in the presence of other languages not only addresses the aforementioned challenge but also holds significance for applications in speech processing and tasks related to natural language processing (NLP). One of the current gaps in speech processing lies in the classification of speakers based on their dialects in the presence of code-mixing. The primary challenge encountered in developing a model for classifying dialects lies in the scarcity of data, particularly when dealing with code-mixed speech. Additionally, accent-based classification is often not feasible for code-mixed data. This is because the presence of multiple languages within a single utterance can obscure the distinctive features of a particular accent. Thus, addressing the scarcity of relevant data and devising effective strategies to handle code-mixed speech are key priorities in the development of dialect classification models.

Ultimately, our objective is to develop a model capable of classifying various dialects within the Telugu language, encompassing both standard and code-mixed speech data. This entails addressing the challenges associated with dialectal variations and code-mixing phenomena.

Since MFCCs (Mel Frequency Cepstral Coefficients) capture phonetic information, we use them for feature extraction in our dialect classification work. This helps us achieve our goal of classifying accents by utilizing the variations in MFCC traits that are present in different languages. A CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory) model is used to classify dialects using these feature vectors as input. In order to capture spatial features inside the MFCCs, the CNN component is very good at understanding local patterns from the data. The model's capacity to identify dialect variants based on speech patterns is improved by the integration of LSTM, which helps it capture temporal dependencies and sequential relations within the retrieved features. For precise accent categorization, this integrated method makes the best use of the temporal and spatial features of the data.

II. LITERATURE REVIEW

Recent research in the field of accent identification and dialect recognition has explored pioneering approaches by utilizing the latest advancements in deep learning and machine learning techniques.Anilkumar et al. [1] focused on developing a Text-to-Speech (TTS) model tailored for the

Indian accent. They employed the Festival Framework with the CLUSTERGEN synthesis method and evaluated the model using the Mel Cepstral Distortion (MCD) metric. While successfully synthesizing Indian-accented speech, the presence of a distortion-like buzz in the synthesized samples indicated a need for further refinement of the synthesis process.Podila and Kommula [2] explored Telugu dialect recognition using deep learning techniques. They achieved promising results with models like LSTM, GRU, BiLSTM, and BiLSTM with Attention Layer, but noted the lack of female speaker representation and suggested incorporating additional speech features beyond MFCCs. Their highest accuracy of 99.10% with the BiLSTM model with Attention Layer demonstrated robust multiclass dialect recognition potential in Telugu.Shamalee Deshpande [3] investigated accent identification between American English and Indian English using formant frequencies of accent markers and Gaussian Mixture Models. While obtaining satisfactory training and testing accuracies for both accents, they suggested improvements in handling misclassification due to low-voiced data frames. Training accuracies reached 76.78% for American accent and 75% for Indian accent, with testing accuracies of 85% and 87.5%, respectively.A supervised autoencoder approach [4] achieved remarkable accuracy for Santali dialect detection, indicating further potential for dataset expansion and evaluation on similar Odia dialects. The methodology achieved 100% accuracy on both development and test sets for Santali dialect detection.Accent identification utilizing artificial neural networks (ANNs) [5] highlighted their potential in accent recognition for speech systems. Despite lacking author attribution, the back-propagation ANN's superior testing accuracy compared to competitive learning suggested its efficacy, achieving 100% testing accuracy.Mannepalli [6] employed Mel Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) for accent recognition in Telugu speech. Despite an overall recognition accuracy of 91%, they suggested improvements in accent-based recognition accuracy and standardization of Telugu speech databases.Methods for identifying languages in Telugu-English code-mixed text [7] achieved high accuracy using models like Naïve Bayes, Random Forest, Hidden Markov Models (HMM), and Conditional Random Fields (CRF). The CRF model achieved the highest accuracy of 91.28%.Gundapu and Mamidi [8] addressed word-level language identification in English-Telugu code-mixed data, showcasing high accuracy with deep learning models. Reported accuracy scores ranged from 91.8091.80% to 98.53%, demonstrating the efficacy of proposed deep learning approaches.Satla and Manchala's research [9] employed a Deep Neural Network (DNN) to identify regional dialects of the Telugu language. Despite dataset scarcity, the DNN model outperformed Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) accuracies, achieving an accuracy of 84.5%.Sahoo's paper [10] proposed a Gaussian Mixture Model (GMM) based method for Mandarin accent identification, achieving promising accent classification error rates. The methodology involved training GMMs independently for each accent and gender, yielding accent classification error rates of approximately 11.7% for females and 15.5% for males.

TABLE I
RESEARCH PAPERS

No.	Title	Author	Dataset	Methodology
1	Identification of Indian English by Speakers of Multiple Native Languages	Ramakrishnan & Vinay Kumar Mittal	Speech Samples particularly Telugu and Kannada	Use of Convolutional Neural Networks (CNNs) for accent identification
2	Telugu Dialect Speech Dataset Creation and Recognition using Deep Learning Techniques	Rama Sai Abhishek Podila & Ganga Sai Sudeep Kommula	Telugu speech recordings representing various dialects (Coastal, Rayalaseema, Telangana)	Recurrent Neural Network (RNN)
3	Accent Classification in Speech	Shamalee Deshpande	Speech samples representing American English and English with an Indian accent	Analysis of formant structures in speech samples
4	Automatic Dialect Detection for Low Resource Santali Language	Sunil Kumar Sahoo	Dataset creation for Santali dialect detection, particularly Santali written in the Odia script	Evaluation of deep learning models for dialect detection
5	Classification of Speech Accents with Neural Networks	Mike V. Chan & Xin Feng	Demographic data and speech features from 22 speakers	Utilization of Neural Network Architectures
6	MFCC-GMM based accent recognition system for Telugu speech signals	Kasiprasad Mannepalli	Telugu speech samples with Coastal Andhra, Rayalaseema, Telangana accents	Extraction of Mel Frequency Cepstral Coefficients (MFCC)
7	WordLevel Language Identification in English Telugu Code Mixed Data	Radhika Mamidi	Cross-script code-mixed languages (Telugu and English)	Exploration of deep learning models like Bidirectional Long Short-Term Memory (BiLSTM)
8	Corpus Creation and Language Identification in Low-Resource Code-Mixed Telugu-English Text	Radhika Mamidi	Cross-script code-mixed languages (English and Telugu)	Investigation of deep learning architectures like Bidirectional Long Short-Term Memory (BiLSTM)
9	Dialect Identification in Telugu Language Speech Utterance Using Modified Features with Deep Neural Network	Shivaprasad Satla1 & Sadanandam Manchal	Speech dataset representing Telangana, Costa Andhra, and Rayalaseema dialects	Application of DNN model with modified features
10	Automatic Accent Identification Using Gaussian Mixture Models	Tao Chen, Chao Huang, Eric Chang and Jingchun Wang	Multi-accent Mandarin corpus consisting of speakers from Beijing, Shanghai, Guangdong, and Taiwan	Training of GMMs independently for each accent and gender

Based on the literature review conducted, it is evident that prior research lacks work specifically focused on dialect classification using code-mixed data, and there is a notable absence of suitable datasets tailored for this purpose. Furthermore, the present frameworks are not able to adequately capture the local features needed for this task, and the lack of approaches to handle low-voiced data can affect the performance of the model. In order to close these gaps and improve model robustness, we are creating a unique dataset that has been carefully selected to reduce low-voiced data. We're introducing a combined framework using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNNs extract local features, and LSTMs capture temporal dependencies in code-mixed speech signals. This integrated CNN-LSTM framework aims to advance dialect classification accuracy in multilingual contexts when code mixed data is present.

III. Materials and Methods

A. Dataset

We have gathered extensive data from popular social media platforms such as Instagram and YouTube, comprising approximately two hours of content for each of the four distinct dialects. Our dataset encompasses diverse age groups, a factor crucial for enhancing the model's performance. Additionally, we have ensured that the dataset includes an equal distribution of male and female speech signals, thereby rendering our data effectively normalized. This comprehensive approach to data collection contributes to the robustness and generalizability of our model.

The four regional variants of Telugu spoken in the Indian states of Andhra Pradesh and Telangana are taken into consideration in our study. Every dialect has unique linguistic traits that produce audible distinctions when spoken. The different cultural and geographical influences present in the region are reflected in these distinctions, which also include variations in pronunciation, vocabulary, grammar, and intonation. Our research focuses on these regional dialects in order to clarify and categorize the distinctive characteristics that set them apart.

B. Methadology

In our dataset, we have four classes representing the four different dialects in the Telugu language. The initial step in developing a recognition system involves extracting features from these audio files. These features are crucial as they serve as the foundation for the subsequent stages of model development. We convert these extracted features into vectors, known as feature vectors, which will be partitioned into training and testing sets for model creation. The quality and relevance of these feature vectors significantly impact the performance and accuracy of our model. Therefore, the method used to generate these features is pivotal in determining the effectiveness of our recognition system.

MFCCs (Mel Frequency Cepstral Coefficients) are widely utilized in tasks such as speaker identification, speech recognition, and even music generation due to their effectiveness in conveying how a person articulates words. We are extracting MFCC features from speech files because they provide valuable insights into speech characteristics. The process of extracting MFCC features involves several steps to optimize model performance. Initially, we apply pre-emphasis to the audio signal, which enhances higher frequencies relative to lower frequencies, improving signal clarity. Next, we segment the audio signal into frames and apply windowing techniques to reduce spectral leakage. Each segmented frame undergoes Fourier transformation to convert it into the frequency domain. Mel filtering is then applied to simulate human auditory responses, capturing relevant frequency bands. The resulting magnitudes are scaled logarithmically to represent them in decibels. Finally, we compute the Discrete Cosine Trans-

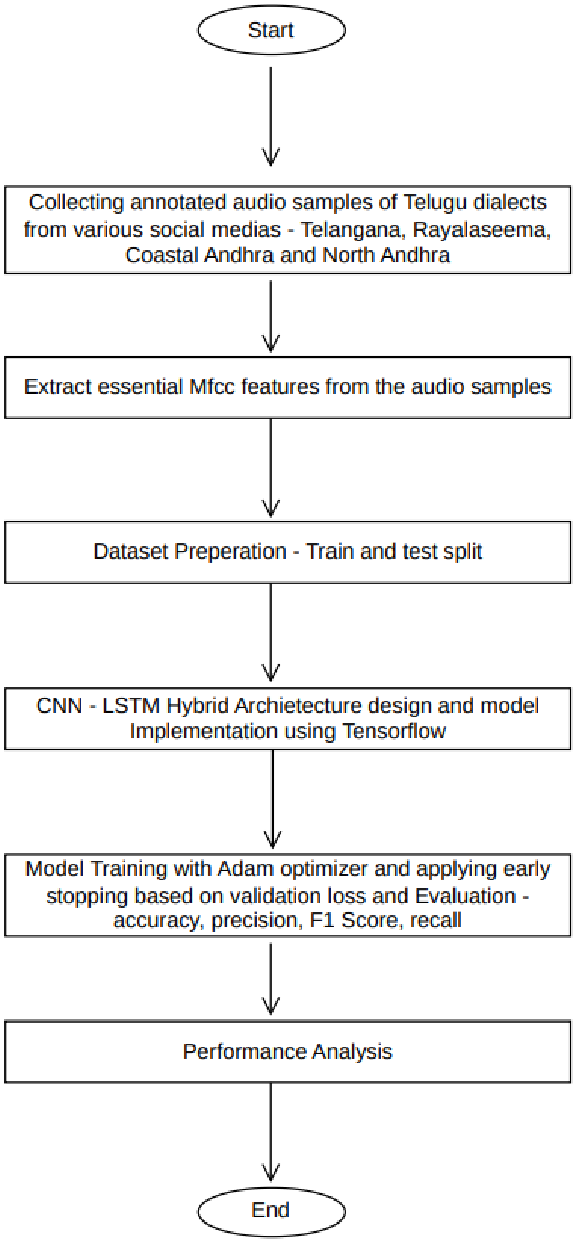
Start
Collecting annotated audio samples of Telugu dialects from various social medias - Telangana, Rayalaseema, Coastal Andhra and North Andhra
Extract essential Mfcc features from the audio samples
Dataset Preperation - Train and test split
CNN - LSTM Hybrid Archietecture design and model Implementation using Tensorflow
Model Training with Adam optimizer and applying early stopping based on validation loss and Evaluation - accuracy, precision, F1 Score, recall
Performance Analysis
End

Fig. 1. OVERVIEW OF THE PROPOSED METHODOLOGY

form (DCT) to extract pertinent coefficients, including overall energy and higher-order features that characterize the spectral properties of the speech signal.

Our CNN-LSTM architecture combines CNNs for spatial feature extraction with LSTM networks for temporal modelling. The CNN component extracted spatial features from MFCCs using convolutional layers and max-pooling procedures, whilst the LSTM component modelled temporal relationships within flattened feature vectors.

For testing, we used a collection of annotated audio samples from Telugu dialects. The MFCC characteristics were extracted from the audio recordings and divided into training and testing sets. The CNN-LSTM hybrid architecture was constructed using the TensorFlow framework, which included three convolutional layers, two LSTM layers, and dropout layers to minimise overfitting.

The Adam optimizer was used for training, with a learning rate of 0.001 over 50 epochs and an early halt due to validation loss. Model performance was assessed using accuracy, precision, recall, and F1-score metrics generated on the test set, which were augmented with confusion matrices for visualisation.

To summarise, our process included extracting features from annotated audio recordings, implementing a CNN-LSTM hybrid architecture, training with the TensorFlow framework, and evaluating performance using multiple metrics. This strategy sought to enhance Telugu dialect recognition by using spatial and temporal interdependence in MFCC characteristics.

IV. Expected Deliverables

A. Trained Model

A completely trained CNN-LSTM hybrid model for Telugu dialect detection. The model should be capable of correctly detecting and categorising various Telugu dialects based on audio samples provided.

B. Evaluation Results:

Comprehensive evaluation results demonstrating the performance of the trained model. Accuracy, precision, recall, and F1-score will be used to evaluate the model's efficacy in dialect recognition.

V.
REFERENCES

[1] R. K. Guntur, R. Krishnan, and V. K. Mittal, “Identification of indian english by speakers of multiple native languages,” in *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2020, pp. 323–336.

[2] R. S. A. Podila, G. S. S. Kommula, K. Ruthvik, S. Vekkot, and D. Gupta, “Telugu dialect speech dataset creation and recognition using deep learning techniques,” in *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 2022, pp. 1–6.

[3] S. Deshpande, S. Chikkerur, and V. Govindaraju, “Accent classification in speech,” in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID’05)*. IEEE, 2005, pp. 139–143.

[4] S. K. Sahoo, B. K. Mishra, S. Parida, S. R. Dash, J. N. Besra, and E. V. Tello, “Automatic dialect detection for low resource santali language,” in *2021 19th OITS International Conference on Information Technology (OCIT)*. IEEE, 2021, pp. 234–238.

[5] M. V. Chan, X. Feng, J. A. Heinen, and R. J. Niederjohn, “Classification of speech accents with neural networks,” in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, vol. 7. IEEE, 1994, pp. 4483–4486.

[6] K. Mannepalli, P. N. Sastry, and M. Suman, “Mfcc-gmm based accent recognition system for telugu speech signals,” *International Journal of Speech Technology*, vol. 19, pp. 87–93, 2016.

[7] S. Gundapu and R. Mamidi, “Word level language identification in english telugu code mixed data,” *arXiv preprint arXiv:2010.04482*, 2020.

[8] S. S. V. Kusampudi, A. Chaluvadi, and R. Mamidi, “Corpus creation and language identification in low-resource code-mixed telugu-english text,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 744–752.

[9] S. Satla and S. Manchala, “Dialect identification in telugu language speech utterance using modified features with deep neural network.” *Traitement du Signal*, vol. 38, no. 6, 2021.

[10] T. Chen, C. Huang, E. Chang, and J. Wang, “Automatic accent identification using gaussian mixture models,” in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU’01.* IEEE, 2001, pp. 343–346.