Automatic Language identification from English-Hindi code-mixed Languages

Akshay Anand
Department of Computer Science
PES University
Bengaluru, India
akshayanand.abbur@gmail.com

Nidhi P G
Department of Computer Science
PES University
Bengaluru, India
nidhigururaj1801@gmail.com

Shreya Ashwath Halgeri
Department of Computer Science
PES University
Bengaluru, India
halgerishreya@gmail.com

Dr. Mamatha H R
Department of Computer Science
PES University
Bengaluru, India
mamathahr@pes.edu

***Abstract*—This paper delves into an advanced technique designed for the automated processing and classification of bilingual audio files that combine Hindi and English speech. With the increasing interconnectedness of the global community, there is a pressing demand for technologies proficient in accurately analyzing and comprehending mixed-language content. Our methodology involves segmenting lengthy audio recordings into shorter, sentence-like units based on silent pauses and utilizing features such as pitch and energy. Subsequently, these segments undergo analysis to ascertain the predominant language—be it Hindi, English, or a blend of both—employing machine learning models. Our classifier training incorporates a diverse set of audio features, notably Mel-frequency cepstral coefficients (MFCCs), pitch tracking, and energy.**

Moreover, our primary objective revolves around the understanding and analysis of English-Hindi code-mixed audio. For this purpose, we harness the capabilities of Long Short-Term Memory (LSTM) networks for training, complemented by the Wav2Vec2FeatureExtractor. This extractor furnishes a numeric representation for the syllables extracted through our segmentation algorithm. Ultimately, our innovative approach culminates in an impressive overall classification accuracy of 71.33 for bilingual audio files.

***Index Terms*—Mel-frequency cepstral coefficients (MFCCs) , pitch , energy,code-mixed audio, Indian Langauges, English-Hindi Code Mixed audio, feature extractor, LSTM,language identification**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Literature Survey

In our literature survey, we meticulously explored various methodologies within the realm of spoken language identification and the detection of languages in code-mixed data, encompassing both textual and audio formats. Additionally, we delved into research on language identification specifically tailored for Indian languages and code-mixed content. By synthesizing insights from these diverse areas, we aimed to inform and enhance our approach to understanding and classifying bilingual audio content blending Hindi and English speech.

[1] deals with classifying different languages- English German, Spanish utilizing audio data. The study uses CNN model where the input is processed as a numpy array and involves several convolutional layers followed by a deep neural network. They used a relatively simple architecture and only a few languages were detected.

In [2], the focus was on the identification of all languages present within mixed spoken language data. This involved leveraging both One-pass and Multi-pass frameworks to address the challenges inherent in multi-lingual speech recognition.To handle English speech recognition, the approach employed the Sphinx model as the primary speech engine. For the Hindi segments of the audio, a transliteration process was employed to convert the spoken Hindi into English text, enabling seamless integration and analysis alongside the English content.To further enhance the accuracy and robustness of the recognition process, a specialized lexicon and phonetic dictionary was developed. These linguistic resources were meticulously constructed using the CMU language toolkit, providing vital support and context for the accurate identification and classification of languages within the mixed-language audio recordings.

Code Mixing Index, a novel metric introduced in [3] was designed to quantify the degree of code-mixing present in textual data.The methodology proposed in [3] focused on word-level identification in code-mixed social media texts.It made used of a dictionary model as a baseline. Techniques such as n-gram pruning and enhanced dictionary models were utilized , with their outputs subsequently ed to an LSTM classifier for final analysis. [4] tackles language identification in English-Telugu code-mixed data using various machine learning models.The models explored include Naive Bayes and Random Forest Classifiers, both enhanced by the use of TF-IDF vectorization techniques, as well as Hidden Markov Models with Viterbi algorithm, and Conditional Random Fields .These latter models are adapted from techniques commonly used in part-of-speech tagging to predict the language of each word within a text.

[6] aimed at identifying English, Kannada, and mixed-language text within a given dataset. This research employs a Bi-directional Long Short-Term Memory model, which is recognized for its ability to capture context from both the past and future input sequences, making it particularly suited for the complexities of language identification tasks. They mention two types of features used in this study: Word Level which is the individual word itself is the primary feature and the model analyzes the letters, spelling and structure of each word to try and identify the language and Predefined Tags. There was an inherent difficulty of analyzing code-mixed languages due to their complex nature, and the presence of sarcastic tags might have further impacted the model's performance.

[7] introduces a methodology designed to detect Sinhala and English words within code-mixed data. This research utilizes sequence tagging as a feature extraction technique, employing character n-grams and the annotations of surrounding words (three words to the left and right of each target word) to enhance the prediction accuracy.The models applied in this study include Support Vector Machines, K-Nearest Neighbors, and XGBoost .

[8] is a study focused on the development of an automatic language identification system tailored for Assamese-English-Hindi data at the word level. This system employs several sophisticated linguistic features for training, including word unigrams, word prefixes and suffixes, tags of previous words, and the context provided by previous and next words. These features are critical for capturing the nuances of language use at the granular level.

[10] is a study that focused on spoken Language Identification (LID) for three Indian languages—Gujarati, Telugu, and Tamil—code-mixed with English. This research employs advanced audio processing techniques, such as Spectral Augmentation and the use of a language mask, to effectively discriminate between language ID pairs in spoken content.The methodology of the study is split into two distinct subtasks to address the challenges of language identification in mixed-language environments. The first subtask involves utterance-level identification to distinguish between monolingual and code-switched utterances. The second subtask is dedicated to frame-level identification, where the specific languages within a code-switched utterance are pinpointed. Temporal masks play a critical role in this process, as they are applied based on the number and positioning of English segments within a code-switched utterance. Both the original and the masked spectrograms are then fed into a sophisticated neural network model.

The model architecture described in the study includes a Convolutional Neural Network (CNN) paired with a Bi-directional Long Short-Term Memory (BiLSTM) encoder network. This setup is particularly adept at capturing long-term sequential context, crucial for accurately identifying languages in mixed speech. The network is trained using the Connectionist Temporal Classification (CTC) loss function, which is well-suited for dealing with alignment and segmentation issues inherent in speech-related tasks. The integration of a dense layer or fully connected (FC) layer preceding the output layer, which employs a softmax function, allows for the generation of a probability distribution over the target labels, providing the final identification of the spoken languages.

The study focuses on developing multilingual and code-switching Automatic Speech Recognition (ASR) systems for seven Indian languages using a hybrid DNN-HMM architecture with the Kaldi toolkit.

Subtask 1 aims to create a multilingual ASR system for six languages (Hindi, Marathi, Odia, Tamil, Telugu, and Gujarati) using a TDNN optimized with MMI, a unified lexicon, and a single 3-gram language model.

Subtask 2 addresses code-switching challenges for Hindi-English and Bengali-English pairs with tailored lexicons and distinct language models for each pair. However, the ASR systems face preprocessing issues like misalignment, inconsistent script usage, punctuation inconsistencies, and mixed words, which could impact model training and accuracy.

In the study reviewed, the aim is to develop sophisticated multilingual and code-switching Automatic Speech Recognition (ASR) systems using the advanced wav2vec 2.0 model for two specific subtasks.

Subtask 1 focuses on creating a multilingual ASR system that supports six Indian languages: Hindi, Marathi, Odia, Telugu, Tamil, and Gujarati. The approach involves training a common multilingual model on a combined character set from all languages, promoting parameter sharing across languages for enhanced generalization. Additionally, the multilingual model functions as a language identifier to select and deploy language-specific monolingual models based on the identified language, optimizing accuracy for each language.

Subtask 2 is aimed at handling code-switching between Hindi-English and Bengali-English pairs. Strategies include developing a common model that incorporates more English data to address the bilingual nature of the input and creating individual models tailored to each language pair to address unique linguistic challenges.

Overall, the study leverages the state-of-the-art wav2vec 2.0 model's capabilities to address the complexities of multilingual and code-switched speech recognition, ensuring robust performance through strategic model training and deployment.

The study develops an Automatic Language Identification system for seven languages—Bengali, Hindi, Telugu, Urdu,

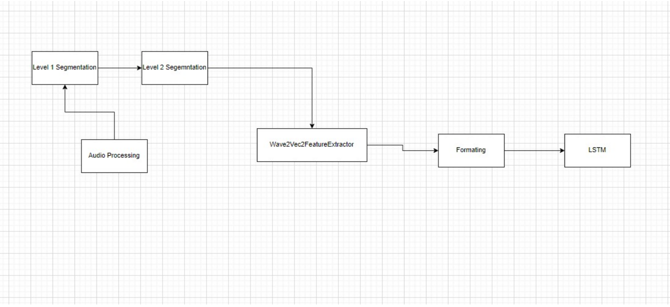

Fig. 1. The block diagram of the proposed methodology

Assamese, Punjabi, and Manipuri—using phonotactic features and prosodic information.

A phonetic engine, trained on speech waveforms and IPA transcriptions, initially focuses on four languages and analyzes speech via syllable segmentation and Short Time Energy contours.

Language identification is conducted using a neural network, specifically an ANN classifier trained for 1000 epochs on a feed-forward back-propagation network with four layers. However, the system faces criticism for its separate model structures, which can vary in prediction accuracy, especially among phonetically similar languages, and the absence of a language model, limiting its decoding and contextual capabilities.

III. PROPOSED METHODOLOGY

A. DATASET

The code mixed audio was obtained from spoken tutorials. These tutorials cover a range of technical topics and the code-switching predominantly arises from the technical content of the lectures . The hindi audio was collected from the IndicTTS database. The english audio was obtained from slidespeech.

B. ARCHITECTURE

1) Audio Segmentation: The selected audio clips underwent two levels of segmentation. In the first level, segments were generated based on silence and low decibel (dB) levels for an extended duration. Following experimentation, optimal values were determined to be a silence threshold of -30dB and a duration of 2 seconds.

Attempts with values such as 0dB, -10dB, and -20dB either failed to produce segments entirely or resulted in truncated words towards the end of segments. To mitigate this issue, an additional 0.5 seconds of silence was appended to the end of each audio segment.

The English and code-mixed audio clips chosen for the experiment were of minute durations, resulting in level one segmentation producing segments measured in seconds. Conversely, the Hindi dataset contained audio clips of seconds duration, leading to either the preservation of the original data or the segmentation of audio into 2-3 segments upon passing through level one segmentation.

MODEL	Accuracy	Precision	Recall	F1
LSTM	0.720	0.710	0.720	0.704

TABLE I
RESULTS

Segments obtained from level one segmentation were subsequently subjected to level two segmentation. In this stage, the audio clips were further decomposed into segments based on their features. Level two segmentation effectively yielded syllabic segments from the provided segments.

2) Feature Representation: The segments generated from the level two segmentation were passed to the Wave2Vec2FeatureExtractor. Wave2Vec2FeatureExtractor is a transformer model designed to provide numerical representations for audio segments.

3) Formatting: The Wav2Vec2FeatureExtractor provided a 3 dimensional tensor for every syllable passed to it. This was converted to 1 dimensional tensor by taking the mean. The length of the tensor representing each syllable was not constant and the number of syllables per level one segment were also not constant. To compensate this, padding was added both at the level one and level two segments, to ensure each tensor was of length 100 and each sentence of 20 syllables length.

4) Model Training: An LSTM model was used for classification. The model had 4 hidden layers and one output layer. The number of units for the hidden layers were 128 , 256, 256, 128, and the output layer had 3 units for classification. The model was trained for 50 epochs. It used Adam optimizer and loss function is sparse categorical cross-entropy.The model was trained for 50 epochs and the best model based on validation accuracy, was chosen for testing.

IV. EXPERIMENTAL RESULTS

An initial LSTM model with only one hidden layer was used, and the results were not satisfactory. While it still achieved an F1 score of 0.6, its performance on code mixed data was not very good. The updated LSTM has a better F1 score and a significant improvement in its performance on code mixed data.

V. ACKNOWLEDGMENT

We express our sincere gratitude to PES University, RR Campus, Bangalore, for the support and resources provided throughout the duration of this project. Special thanks are extended to the Faculty of the Department of Computer Science and Engineering for their invaluable guidance. We are particularly grateful to Dr. Mamatha H.R. and Dr.Preethi P, our Natural Language Processing course faculties, whose insights and expertise greatly enhanced our understanding and enriched our work on this project.

REFERENCES

[1] S. Mukherjee, N. Shivam, A. Gangwal, L. Khaitan and A. J. Das, ”Spoken Language Recognition Using CNN,” 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2019

[2] Bhuvanagiri, K., and Sunil Kopparapu. ”An approach to mixed language automatic speech recognition.” Oriental COCOSDA, Kathmandu, Nepal (2010).

[3] Das,Amitava and Gambäck, Björn “Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text”,International Institute of Information Technology, 2014

[4] Gundapu, Sunil And Mamidi, Radhika “Word Level Language Identification in English Telugu Code Mixed” Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong 2018

[5] C. Madhu, A. George and L. Mary, ”Automatic language identification for seven Indian languages using higher level features,” 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kollam, India, 2017, pp. 1-6, doi: 10.1109/SPICES.2017.8091332. keywords: Speech;Feature extraction;Speech recognition;Hidden Markov models;Stress;Engines;Rhythm;Language Identification;Phonotactics;Phonetic Engine;Prosody;Multilayer feedforward neural network,

[6] Rangan, Pradeep , Teki, Sundeep Misra, Hemant. (2020). Exploiting Spectral Augmentation for Code-Switched Spoken Language Identification.INTERSPEECH-2020 - ”First Workshop on Speech Technologies for Code-switching in Multilingual Communities 2020

[7] Diwan, A., Vaideeswaran, R., Shah, S., Singh, A., Raghavan, S., Khare, S., Unni, V., Vyas, S., Rajpuria, A., Yarra, C., Mittal, A., Ghosh, P.K., Jyothi, P., Bali, K., Seshadri, V., Sitaram, S., Bharadwaj, S., Nanavati, J., Nanavati, R., Sankaranarayanan, K. (2021) MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. Proc. Interspeech 2021, 2446-2450, doi: 10.21437/Interspeech.2021-1339

[8] Chadha, H. S., Shah, P., Dhuriya, A., Chhimwal, N., Gupta, A., Raghavan, V. (2022). Code Switched and Code Mixed Speech Recognition for Indic languages. arXiv preprint arXiv:2203.16578.

[9] Ali A, Dehak N, Cardinal P, Khurana S, Yella SH, Glass J, Bell P, Renals S. Automatic dialect detection in arabic broadcast speech. arXiv preprint arXiv:1509.06928. 2015 Sep 23.

[10] Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word Level Language Identification in Code-mixed Kannada-English Texts using Deep Learning Approach. In Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, pages 29–33, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

[11] I. Smith and U. Thayasivam, ”Language Detection in Sinhala-English Code-mixed Data,” 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 2019, pp. 228-233, doi: 10.1109/IALP48816.2019.9037680. keywords: Hidden Markov models;Data models;Facebook;Task analysis;Support vector machines;Encoding;code-mixed;code-switching;Sinhala-English;language detection;social media data,

[12] Bora MJ, Kumar R. Automatic word-level identification of language in assamese english hindi code-mixed data. In4th workshop on indian language data and Resources, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) 2018 May 12 (pp. 7-12).