**MANGALORE  UNIVERSITY**

**Title of the project**

**“SPOKEN LANGUAGE IDENTIFICATION”**

**BY**

**THRIVENI N**

**Under the Guidance of**

**Dr. H.L. Shashirekha Professor**

**Department of Computer Science**

**Mangalore University**

**Mangalagangothri-574199**

**DEPARTMENT OF POST GRADUATE STUDIES AND RESEARCH IN COMPUTER SCIENCE**

**MANGALORE UNIVERSITY**

**MANGALAGANGOTHRI-574199, KARNATAKA, INDIA**

**Declaration**

This is to declare the project entitled “**Spoken Language Identification**”is a bonafede work independently carried out by me, **Thriveni N**, Student of 3rd semester Masters in Computer Applications under the supervision and guidance of **Dr. H. L. Shashirekha**, Professor, Department of Post-Graduate Studies and Research in Computer Science, Mangalore University. This is submitted in partial fulfillment of the requirement for the award of Masters in Computer Applications Further, it is declared that this project is the result of my own efforts and has not been submitted to any other university for the award of any other Degree or Diploma.

Place: Mangalagangothri Thriveni N

Date: 14-03-2023 Reg. No.: P05AZ21S0194

**Abstract**

 Spoken Language Identification (SLID) is recognizing the language being a talk by an anonymous speaker from an audio clip. Humans are the most error-free language identification system. Spoken Language Identification is the process of detecting the language of an utterance by an anonymous speaker, irrespective of gender, accent and pronunciations. Implementation of an acoustic model for Spoken Language Identification is to be carried out in this project. The major task is to identify those features or parameters which could be used to clearly distinguish between languages. This acoustic model makes use of mean values of Mel Frequency Cepstral Coefficients (MFCC). The system uses k-Nearest Neighbor (kNN) model, Decision Tree, Logistic Regression and Random Forest algorithms to the handle the problem of multi class classification. The project aims at detecting Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu.

Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio clips. This corpus comprises of utterances, each of them spanning over a uniform duration of 4 to 7 seconds. The entire corpus is split into two sets, larger unit as the training dataset and a smaller set as the test set. Preliminary results indicate an overall accuracy of 97%, 97%, 100% and 100% respectively. Thus, the acoustic model employing mean values of MFCC proves to be a viable approach for Language Identification.

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Contents** | **Pg.No** |
| 1. | Introduction |  |
| 2. | Literature Survey |  |
| 3. | Methodology |  |
| 4. | Experiments and results |  |
| 5. | Conclusion |  |
| 6. | References |  |

1. **Introduction**

Language Identification (LID) is the process of determining and classifying a natural spoken language from given content and datasets. It is undertaken by performing computational linguistics approaches and applying many contexts. These contexts include; text categorization of a written text or speech recognition of a recorded utterance of a spoken identified language. It is a challenging task because due to the variations in the type of speech input and understanding how humans process and interpret speech in adverse conditions.

When using a LID system, several types of information are considered. Furthermore, human understanding has inspired the classification of information, and several studies have applied methods which people have used to differentiate languages, whether consciously or not. A broad classification has been used to separate or split speech features into a low level and a high level. The acoustic features usually modeled by MFCCs are the compact representation of the input speech signal fulfilling a compression of the data contained in the audio waveform.

LID is an important pre-processing technique applied to future multi-lingual speech processing systems, such as audio and video information retrieval, automatic machine translation, multi-lingual speech recognition, intelligent surveillance and etc. A major problem in LID is how to design a specific and effective language to represent speech utterances. The features of the kNN model, Decision Tree model, Logistic Regression and Random Forest models are used as input to the classifier with the authors emphasizing the high capabilities of all mentioned based features in providing a discriminative factor for classifying languages.

1. **Literature Survey**
   1. Montavon, Gregoire. "Deep learning for spoken language identification." In *NIPS Workshop on deep learning for speech recognition and related applications*, pp. 1-4. 2009.

Spoken language identification is the problem of mapping continuous speech to the language it corresponds to. Applications of spoken language identification include front-ends for multilingual speech recognition systems, web information retrieval, automatic customer routing in call centers or monitoring. Empirical results have shown that many systems based on the manual extraction of acoustic, prosodic, phonotactic or lexical features have significantly lower performance on small speech samples than on large speech samples, while a human would still be successful. A hypothesis for this low performance is that the set of extracted features is insufficient. Deep learning potentially addresses this issue by exploring the space of features automatically, bypassing the traditional phoneme recognition layer and learning instead purely discriminative features. A deep architecture is implemented and evaluated on several datasets.

* 1. Shukla, Shikhar, and Govind Mittal. "Spoken language identification using convnets." In *Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15*, pp. 252-265. Springer International Publishing, 2019.

Language Identification (LI) is an important first step in several speech processing systems. With a growing number of voice-based assistants, speech LI has emerged as a widely researched field. To approach the problem of identifying languages, we can either adopt an implicit approach where only the speech for a language is present or an explicit one where text is available with its corresponding transcript. This paper focuses on an implicit approach due to the absence of transcriptive data. This paper benchmarks existing models and proposes a new attention based model for language identification which uses Log-Mel spectrogram images as input. We also present the effectiveness of raw waveforms as features to neural network models for LI tasks. For training and evaluation of models, we classified six languages (English, French, German, Spanish, Russian and Italian) with an accuracy of 95.4% and four languages (English, French, German, Spanish) with an accuracy of 96.3% obtained from the VoxForge dataset. This approach can further be scaled to incorporate more languages.

* 1. Aarti, Bakshi, and Sunil Kumar Kopparapu. "Spoken Indian language identification: a review of features and databases." *Sādhanā* 43 (2018): 1-14.

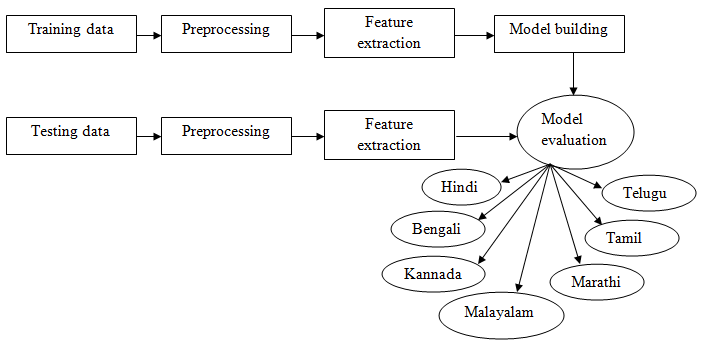
Spoken language is the most natural mode of communication in today’s world, especially given the advances that have happened in the area of automatic speech recognition (ASR). However, to achieve good ASR performance in terms of recognition accuracies, two things are crucial, namely (a) the correct identification of the spoken language, which in turn depends on (b) the availability of a good speech corpus. From the Indian context, there has not been much work done in either of these two crucial areas, especially given that India boasts of many languages.

* 1. Draghici, Alexandra, Jakob Abeßer, and Hanna Lukashevich. "A study on spoken language identification using deep neural networks." In *Proceedings of the 15th International Audio Mostly Conference*, pp. 253-256. 2020.

Spoken language identification deals with classifying the language associated to a speech recording. Possible application scenarios involve user interfaces for information retrieval systems as well as systems for automatic speech recognition, digital law enforcement, multilingual translation systems, emergency call routing, and spoken document retrieval. The main challenge of spoken language identification is find meaningful audio feature representations which are robust to individual variations in pronunciation as well as to similarities of languages within the same language families.

1. **Methodology**

**3.1 Framework:**



**Figure 1: Framework**

**As shown in the Figure 1: Framework, we have done several steps to classify the languages. The training data and testing data are taken as input. This speech dataset includes languages Hindi, Bengali, Kannada, Malayalam, Marathi, Tamil and Telugu. Preprocessing steps are worked for both training and testing dataset. Using MFCC technique, feature extraction has been done. Model was builded for training data.** kNN, Random forest, Logistic regression and Decision tree algorithms are used to build the model. **After building model, the model was evaluated. This model classifies the languages as Hindi, Bengali, Kannada, Malayalam, Marathi, Tamil and Telugu.**

**3.2 Sampling rate:**

Sample rate is the number of amplitudes present in the frequency. In signal processing, sampling is the reduction of a continuous signal into a series of discrete values. The sampling frequency or rate is the number of samples taken over some fixed amount of time. Look at the sampling rate of the audio signals.

**3.3 Split into train and test set:**

Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio clips. This corpus comprises of utterances, each of them spanning over a uniform duration of 4 to 7 seconds. The entire corpus is split into two sets, larger unit as the training dataset and a smaller set as the test set. Trained the model on 70% of the data and validate on the remaining 30%.

**3.4 Preprocessing the audio waves:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first step while creating a machine learning model. In the data exploration part earlier, we have seen that the duration of a few recordings are different. So we used to convert the time domain data into frequency domain data. By this technique the frequency is taken easily for modeling.

**3.5 Feature Extraction technique:**

**MFCC (Mel Frequency Cepstral Coefficients):** MFCC is a feature extraction technique used in machine learning. MFCC of a signal are a small set of features which concisely describe the overall shape of spectral envelope. They are derived from a type of cepstral representation of the audio clip.

**3.6 Model building algorithms:**

By extracting features of training data and testing data, model will build. Here we used models are kNN, Decision Tree, Logistic Regression and Random Forest. Machine learning models use a set of input values to predict output values.

**kNN (k-Nearest Neighbor):** kNN algorithm, it will work the most frequent label in the case of classification and averages the labels in the case of regression. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

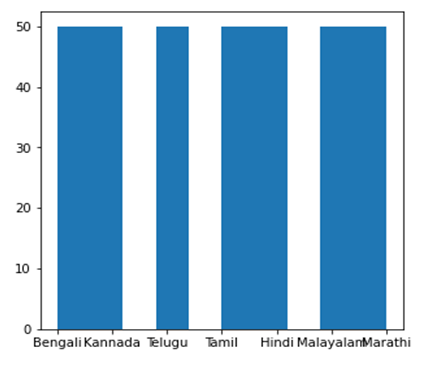
**Decision tree:** Decision tree algorithm is a supervised learning technique that can be used for both classification and regression problems. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. A decision tree is a graph that uses a branching method to illustrate every possible output for a specific input.

**Logistic regression:** Logistic regression is machine learning algorithm, which comes under the supervised learning technique. Logistic regression predicts the output of a categorical dependent variable. Logistic regression is easier to implement, interpret, and very efficient to train. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

**Random Forest:** Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset takes the average to improve the predictive accuracy of that dataset. Random forest algorithm builds a forest in the form of an ensemble of decision trees which adds more randomness while growing the trees. While splitting a node, the algorithm searches for the best features from the random subset of features which adds more diversity, thereby resulting in a better model.

* 1. **Dataset:**

A dataset is a structured collection of data generally associated with a unique body of work. In this project, the dataset is taken from **OpenSLR**, A multilingual corpus of talks for speech recognition and translation. Here we use different languages with 350 number of voice recordings. Each file has a size of 4 to 7 seconds. Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio clips. Mainly 7 Indian languages are taken. In this spoken language identification labels are settled as Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. Here below we can see the graphical representation of dataset -



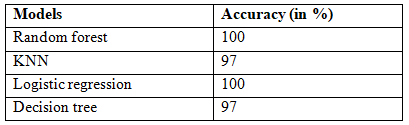
**3.8 Classification:**

It is the problem of identifying which of a set of languages an observation belongs to. A supervised machine learning method where the model tries to predict the correct label of a given input data. By using different model algorithms classification and regressions are done. The current system is capable of identifying Bengali, Kannada, Hindi, Malayalam, Marathi, Tamil and Telugu with an appreciable accuracy.

1. **Experiments and Results**

**Experiments:** At first, some pre-processing is carried out by converting time domain to frequency domain. kNN, Random forest, Logistic regression and Decision tree algorithms are used to build the model. Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio clips. This corpus comprises of utterances, each of them spanning over a uniform duration of 4 to 7 seconds. The entire corpus is split into two sets, larger unit as the training dataset and a smaller set as the test set. Preliminary results indicate an overall accuracy of 97%, 97%, 100% and 100% respectively as shown in the results. Mainly 7 Indian languages are taken. In this SLID labels are settled as Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu.

**Results:** The accuracy of the performance is mentioned below -



1. **Conclusion**

Spoken Language Identification is a challenging problem to deal with. The performance of language identification System is mainly depending on the quality of Signal Preprocessing Stage. Powerful performance can be achieved using relatively short files with minimum preprocessing. The Preprocessing quality is giving the biggest impact on the Language Classification performance. An Improvement in any individual part can improve the overall system performance.

**6. References**

* 1. Montavon, Gregoire. "Deep learning for spoken language identification." In *NIPS Workshop on deep learning for speech recognition and related applications*, pp. 1-4. 2009.
  2. Shukla, Shikhar, and Govind Mittal. "Spoken language identification using convnets." In *Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15*, pp. 252-265. Springer International Publishing, 2019.
  3. Aarti, Bakshi, and Sunil Kumar Kopparapu. "Spoken Indian language identification: a review of features and databases." *Sādhanā* 43 (2018): 1-14.
  4. Draghici, Alexandra, Jakob Abeßer, and Hanna Lukashevich. "A study on spoken language identification using deep neural networks." In *Proceedings of the 15th International Audio Mostly Conference*, pp. 253-256. 2020.