International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

Spoken Language Detection Using Deep Learning

G.V. Satyanarayana[1], D.Varun[2], D.Sandeepika[3], B.Uma Devi[4], B.Rohit Kumar[5]

Raghu Institute of Technology, Visakhapatnam, Andhra Pradesh, India.

Submitted: 25-05-2022 Revised: 01-06-2022 Accepted: 05-06-2022

ABSTRACT

The task of recognizing the speech of an audio clip by an unknown speaker, regardless of the speaker's gender, age, and speaking style is called spoken language identification. The approach reflects a scenario from reality found in modern surveillance/communication organizations where semi-automated systems indexing/documentation is implemented, which could be facilitated by the proposed speech recognition pre-processing. The formidable task is to recognize the features that can distinguish between languages clearly and efficiently. The model takes audio files and converts these files into spectrogram images. It applies the Convolutional Neural Network (CNN) to highlight the main attributes or characteristics of the and automatically recognize speech. The main aim is to detect languages between English, Telugu, Hindi, and Tamil. An experiment was performed on different audio files using the VoxLingua107 dataset [1]. We generate semi-random search terms from language-specific Wikipedia data, which are then used to retrieve videos from YouTube for 107 languages. Speech activity detection and speaker polarization are used to extract segments from the video containing the voice. Post-filtering is used to remove segments from the database that are not available in certain languages, increasing the percentage of correctly labelled segments to 98%, based on verification of various assets. The size of the resulting training set (VoxLingua107) was 6628 hours (average 62 hours per language) and it was accompanied by an evaluation set of 1609 validated statements [1]. We use the data to build a language recognition model for some speech recognition tasks from videos containing voices.

The whole data set is divided into training and test data sets. The preliminary results give an overall accuracy of 98%. Extensive and precise testing shows an overall accuracy of 90%.

KEYWORDS: Convolutional Neural Network (CNN), VoxLingua107.

I. INTRODUCTION

Spoken Language Detection (SLD) is the task of automatically classifying spoken language from the given utterance of audio [1]. People all over the world have been united by the phenomenon of globalization. However, one obstacle to this increase in global communication is that various languages are used by people around the globe and effectively we do not have a common way to communicate with everyone. That is, in order to communicate successfully, language that is understandable mutually to both parties is required. Language detection offers a means that provides flexibility to erase the communication barrier. Speech can be in spoken or text form. Spoken language detection is the task of recognizing spoken language. The task of identifying the spoken language from a speaker's speech sample is stated as Automatic language detection/identification. Humans are the most accurate speech recognizers in this current world. They can recognize any language within a couple of seconds of hearing whether they understand that language or don't have an idea about that language. They can also judge whether it's similar to any language that they are familiar with. Each expression is nothing more than an audio signal or a language.

The audio signals are processed by Speech processing which is the study of signals. Signals are usually processed in a digital representation, so speech processing can be considered a special case of digital signal processing. Properties of a language can be showcased by various aspects embedded in that language. The need for a good interface arises as the complex raw signal may not be good enough to feed the voice detection system as input.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

SLD is used as a pre-processing step in various applications such as automatic call transfer, multilingual language translation and human-machine communication systems, multilingual voice transcription systems, and voice document retrieval. SLD is also widely used in the intelligence and security field.

The objective of this work was to investigate whether automatically extracted and labelled voice data from the web could be used to build SLD systems. Our aim is to target different broadband audio conditions and provide extensive language coverage. We extract audio data from retrieved YouTube videos using random language-specific search terms. If the language of the video title and description matches the language of the search term, the video audio may be in that particular language. This makes it possible to gather huge amounts of slightly noisy data at a relatively low cost. In this work, we designed a system that allows the user to record audio and detects the spoken language spoken by the user. In this work, we used the VoxLingua107 dataset which consists of 6628 hours (average 62 hours per language) and it was accompanied by an evaluation set of 1609 validated statements [1]. We used the data from English, Telugu, Hindi, and Tamil languages to train the model and detect the audio from one of these languages. We used Convolutional Neural Network (CNN) algorithm to detect the language,CNN is a special kind of recurrent neural network capable of handling long-term dependencies. This model is analyzed with different deep learning and machine learning techniques. On various evaluation metrics, the proposed approach differs from various state-of-the-art methods and shows the comparison with different techniques.

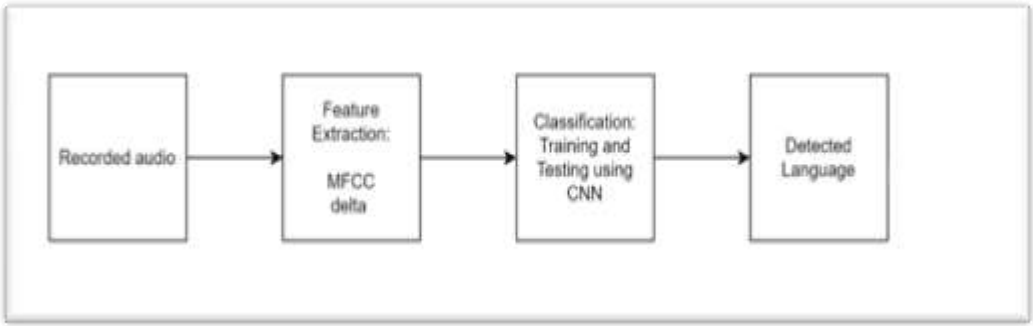


Figure 1.1 Schematic diagram of spoken language detection

II. LITERATURE REVIEW

In the early 1970s the research in the field of Spoken Language Identification commenced. Almost 5 decades of research, various methods in various aspects were studied to achieve a high-performance language recognition system. Spoken language detection is one of the several tasks in which information is extracted from a speech signal and that information is used to detect the spoken language. The information that is drawn out from this information can be of any form phonotactic, acoustic, or prosodic. The phonotactic approach is the process that deals with modelling speech at the syllable level or phoneme. A phoneme is a sound or a group of sounds that is the smallest unit that can be used to differentiate between utterances. Various approaches that are based on the Phoneme based features have been proposed in the field of LiD [2].

A task-independent spoken language identification which uses a Large Vocabulary Automatic Speech Recognition (LVASR) was proposed by Hieronymous and Kadambe [3]. The LVASR LiD system has many differences in the language model. Different languages have different numbers of phonemes, different word lengths, and different word sequences which may contain high-frequency words.Berkling and Barnard proposed a Broad Phonema [4] method for identification of Spoken language. Their system discriminated between English and Japanese with 90% accuracy. The duo also proposed a theoretical error prediction for a language identificationsystem [5]. Based on the presumption that the acoustic structure of language can be estimated by segmenting the speech into phonetic categories a segmental approach to Automatic Language Detection is designed [6].

The performance of four approaches [7] for automatic language identification of speech utterances was compared by Zissman, single-language phone recognition which was followed by language-dependent, interpolated n-gram language modelling (PRLM); parallel PRLM, which uses multiple single-language phone recognizers, each trained in a different language; and language-dependent parallel phone recognition (PPR). Spoken Language Identification using Machine Learning.Another path taken for spoken language

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

identification is the prosodic approach. A large number of vocal tract dependent features like pitch, rhythm, and stress are encompassed from prosodic features. An approach using pitch contour information for automatic language identification is proposed by Lin and Wang [8].

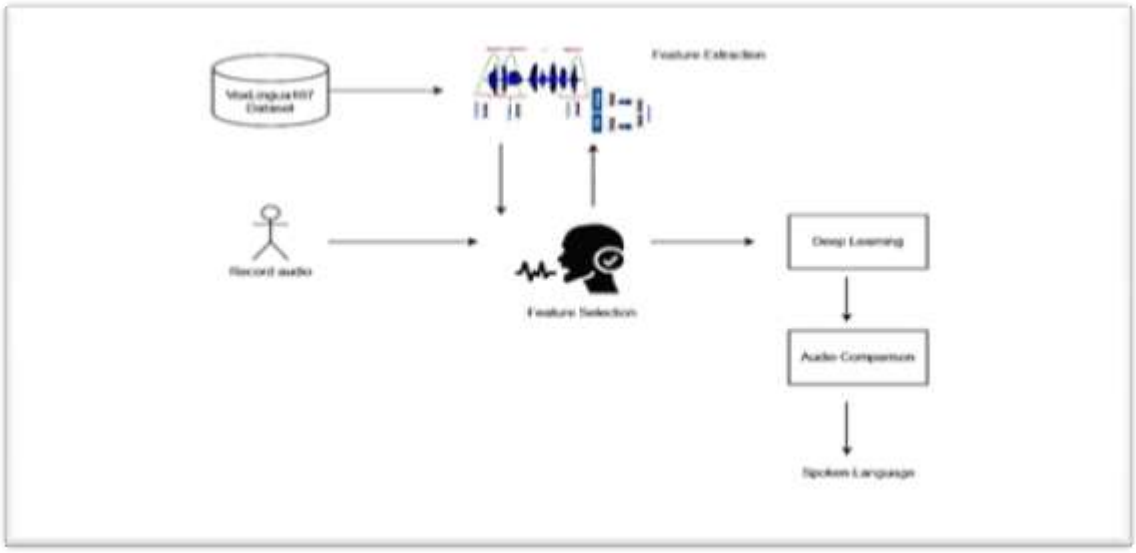
A section of prosody is approximated by a group of Legendre polynomials so coefficients of polynomials kind a feature vector to represent this pitch contour. Brady and Hirschberg [9] examined the role of intonation and rhythm across four Arabic non-standard speeches: Gulf, Iraqi, Levantine, and Egyptian for the aim of automatic dialect recognition. Sensible results with the period of utterances being 2 minutes were given by this technique.

III. EXISTING SYSTEM

Various experiments are planned to work out the issue of developing an Automatic Language Detection system using various approaches like acoustic approach. Many of them tried using Generative Adversarial Networks (GANs) for language identification for hardiness on semi-supervised, unsupervised tasks. They used Support Vector Machine(SVM) classifiers for identifying spoken language from speech utterances but there are many flaws in using SVM.

SVM doesn't work well on short utterances, giving less accuracy. Standard identification systems are supported by i-vector systems for spoken language process tasks, that are inefficient.

3.1. SYSTEM ARCHITECTURE



3.1 system architecture

IV. PROPOSED SPOKEN LANGUAGE DETECTION FRAMEWORK

This session discusses the spoken language detection framework

4.1. OBJECTIVE

Due to globalization need for communication with various people across the globe has been an obstacle as there is no common language to communicate across the globe and most people are not familiar with all languages. It is hard to recognize language spoken by others if we are unaware of that language. Identifying language spoken by different people of different age groups, gender, places, and assents is an obstacle. To identify spoken language using an application is also quite difficult as there are many obstacles like background noise, improper voice, etc. A deep learning CNN technique is proposed to extract features. Without automatic language detection, speech cannot be parsed correctly and grammar rules cannot be applied, causing subsequent speech recognition steps to fail. Current Automatic Speech Recognition (ASR) systems require the user to manually specify the correct system input language for proper operation. The goal of this work is to develop a system that automatically detects the language a user speaks among English, Telugu, and Hindi and Tamillanguages.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

4.2. Proposed spoken language detection framework

A prediction is made based on a deep learning model that can classify the various languages accurately. A novel deep learning-based CNN classifier is used to draw out attributes from images. The proposed model is analyzed with various machine learning and deep learning techniques over 8 languages. In the proposed system, we used Convolution Neural Network (CNN) classifier which has the ability to memorize data which helps in identifying languages more accurately.

4.3. PREPROCESSING

In the preprocessing phase, we pre-process the data such that we get rid of flaws in the data. While working with audio files the main obstacle is background noise,data is prepossessed, to get accurate and clean data. In this model, we use Mel-frequency cepstrum (MFCC) and delta to extract all important features from the data i.e speech utterances or audio signals.

4.4. DATASET DESCRIPTION

We used the VoxLingua107 [1] dataset in our model this dataset consists of audio files from 107 different languages that are extracted from the YouTube videos using semi-random search phrases from language-specific data from Wikipedia. There is a total of 107 different language audio files among which we used data for 4 languages that are English, Telugu, Hindi and Tamil. The size of the resulting training set is 3200 seconds (400 seconds per language on an average) and it is accompanied by an evaluation set of 800 verified utterances. Each audio has an exact duration of 10 seconds (sharp) with a sample rate of 22050, a bit depth of 16 bits, and channels 1. We use this data to develop a model for recognizing spoken language spoken by the user among these languages. Many experiments showed that using automatically retrieved training data gives competitive results to using hand-labeled proprietary datasets. This dataset is available publicly to everyone [10].

4.5. MODEL DESCRIPTION

The version description describes the framework for all of the fashions for the test purpose.

- A Convolutional layer is usually accompanied via way of means of the right pooling layer which facilitates to encompass the explosion of attributes and maintains the version efficiency and keeps it small and easy.

Each layer in CNN is likewise accompanied via way of means of a dropout layer, Rectified Linear Units (ReLU), and batch normalization.

- The batch normalization is chargeable for the convergence of learned representations.
- A dense layer is used at the end which is very useful to classify output as it acts as an output layer of the designed version.

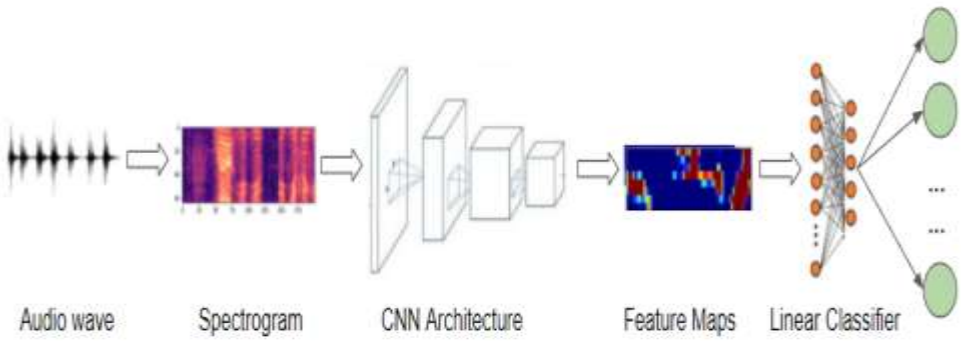


Figure 4.5.1 CNN Architecture

4.6. MODEL CHARACTERISTICS

Table 4.6.1 represents the detailed description of the model used in this system which also shows various hyper parameters

4.7. MODEL DETAILS

This approach uses the Convolutional Neural Network (CNN), a machine learning technique that identifies language from the dataset given. All data is split into training and testing data, using the training data to train the model and the testing data to test the output. This CNN approach gave 97% accuracy.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

TABLE 4.6.1 DETAILED DESCRIPTION OF THE MODEL

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 37, 860, 32)       320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 18, 430, 32)       0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 16, 428, 64)       18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 8, 214, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 6, 212, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 3, 106, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 1, 104, 256)       295168
_________________________________________________________________
flatten (Flatten)            (None, 26624)             0
_________________________________________________________________
dense_16 (Dense)             (None, 9)                 239625
=================================================================
Total params: 627,465
Trainable params: 627,465
Non-trainable params: 0
```

V. RESULT AND DISCUSSION

This section discusses about the result of the model. In this work we tested the data using CNN model which gave an accuracy of 96% whereas while tested using Long Short-Term Memory model we got an accuracy of 86%.

5.1. PERFORMANCE OF CLASSIFICATION MODEL: CONFUSION MATRIX

In this section we see a confusion matrix drawn from the CNN model which gives the information about the performance of the model as shown in the Figure 5.1.1 Multiclass classification represents eight languages English, Telugu, Hindi and Tamil. In this matrix, all the diagonal values are the correctly predicted values which are called true values while others are false values that are predicted incorrectly.

In this matrix x-axis is the true labels whereas y-axis contains predicted labels and all the diagonal values are true positive values that are predicted correctly.

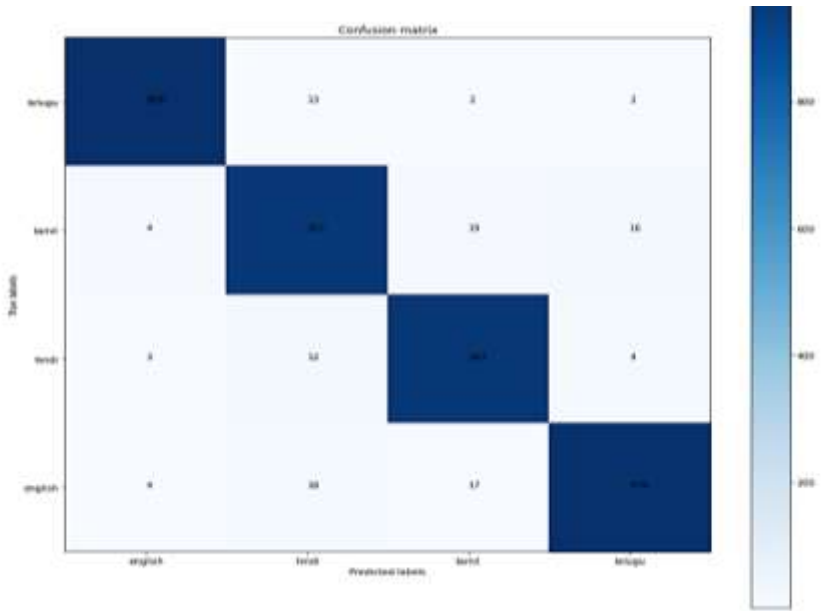


Figure 5.1.1 Multiclass classification

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

5.2. PERFORMANCE EVALUATION

In the Figure 5.2.1 we can find various metrics of our model like F1- Score, Precision, Recall and Support. The formula for calculating these metrics are

Precision = tp / (tp + fp)
Recall = tp / (tp + fn)
F1 Score = 2 * (Precision*Recall)/ (Precision+Recall)

Where truly predicted values are called as true positive (tp) values where predicted output value matches the original output values and incorrect predicted values where predicted output values did not match with original output values. In our model we have 4 languages and for these languages the precision, recall and f1-score and accuracy compared with various other algorithms such as LSTM and Feed Forward Neural Network are compared and shown in the figure 5.2.1.

LANGUAGE	MODEL	PRECISION	RECALL	F1-SCORE	ACCURACY
ENGLISH	FFNN	93	92	92	90
	LSTM	88	88	88	86
	CNN	98	99	99	97
HINDI	FFNN	89	89	89	90
	LSTM	87	86	87	86
	CNN	96	96	96	97
TAMIL	FFNN	90	92	91	90
	LSTM	86	87	86	86
	CNN	98	96	97	97
TELUGU	FFNN	91	90	90	90
	LSTM	86	86	86	86
	CNN	96	98	97	97

Figure 5.2.1 Classification report of various models

A receiver operating characteristic (ROC) curve is a graph that represents the classification model at different classification threshold values. The two parameters or attributes of ROC are false-positive rate (FPR) and true positive rate (TPR) are plotted using this ROC curve.

In figure 5.2.2 a multiclass ROC curve for language identification is presented. This spoken language detection contains four languages: English, Telugu, Hindi, and Tamil.

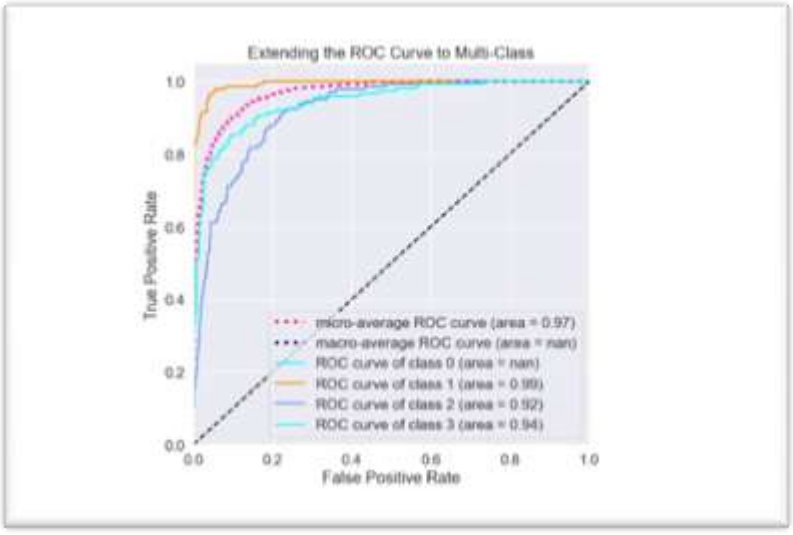


Figure 5.2.2 a multiclass ROC curve

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 113-119 www.ijaem.net ISSN: 2395-5252

VI. CONCLUSION

The current system is able to identify English, Hindi, Telugu and Tamil with remarkable accuracy. Efforts are being made to adapt the LiD for regional languages such as Telugu, Kannada, and Hindi. The main obstacle in any LiD research is the availability of a standard corpus of multilingual language for training. This project did not use standard datasets, but still competes for good accuracy. We used a new dataset VoxLingua107 and experiments on the NIST LRE07 evaluation data showed that using VoxLinugua107 data results in a classifier that is not far in accuracy from a model trained on large amounts of in-domain data.

REFERENCES

[1]. http://bark.phon.ioc.ee/voxlingua107/
[2]. https://www.arxiv-vanity.com/papers/2011.12998/
[3]. K. M. Berkling, T. Arai and E. Barnard, "Analysis of phoneme-based features for language identification", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Adelaide, Australia, April 1994
[4]. J. Hieronymous and S. Kadambe, "Spoken Language Identification Using Large Vocabulary Speech Recognition", in Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, 1996.
[5]. K. M. Berkling and E. Barnard, "Language Identification of Six Languages Based on a Common Set of Broad Phonemes", in Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 1994.
[6]. K. M. Berkling and E. Barnard, "Theoretical Error Prediction for a Language Identification System using Optimal Phoneme Clustering", in Proceedings 4rd European Conference on Speech Communication and Technology (Eurospeech 95), Madrid, Spain, September 1995.
[7]. Y. K. Muthusamy, "A Segmental Approach to Automatic Language Identification", Ph.D thesis, Oregon Graduate Institute of Science & Technology, July 1993.
[8]. M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in IEEE Trans. Speech and Audio Proc., SAP-4(1), January 1996.
[9]. Chi-Yueh Lin, Hsiao-Chuan Wang, “Language identification using pitch contour information”, from Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwa
[10]. Fadi Biadsy, Julia Hirschberg, “Using Prosody and Phonotactics in Arabic Dialect Identification”, Department of Computer Science, Columbia University, New York, NY, 10027