IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LID SYSTEMS FOR SPEECH PROCESSING SYSTEMS

(GMM based Bayes classifier model, FCN and CNN models to automatically recognize and classify the spoken language)

[1]Hamsa A S, [2]Sushma V

[1]Assistant Professor, [2]Assistant Professor
[1]Department of Computer Science & Engineering,
[1]ATME College of Engineering, Mysore, Karnataka, India

Abstract: The Language Identification Systems should identify the language irrespective of gender, accents and pronunciations. Automatic language identification has always been a challenging issue and an important research area in speech signal processing. The use of acoustic model to get only those features which are independent of prosodic or phonotactic information are used to model languages. The language considered here includes set of 4 – Indian Languages. An Acoustic model is used to extract suitable features from speech samples across different languages and different speakers in each language. It is the ability of the Deep Neural Network's technique to perform complex correlation among speech signal features, which enhances its performance over traditional approaches, here different DNN models like GMM, FCN and CNN's are implemented for an acoustic method like MFCC.

Index Terms **–MFCC, Acoustic model, Prosodic\Phonotactic, Mel Scale.**

I. Introduction

Language as a communication system is thought to be fundamentally different from and of much higher complexity than those of other species as it is based on a complex system of rules relating symbols to their meanings, resulting in an indefinite number of possible innovative utterances from a finite number of elements. Among the various factors that define different cultures and communities, an important factor is language. The importance of speech and language for human to human communication can be over emphasized. Speech would thus be the most natural medium of interaction between humans and machines too. Language can be in the spoken or textual form. Spoken Language Identification (LID) is the process of identification of the language spoken in an utterance.

Automatic language identification is the problem of identifying the language being spoken from a sample of speech by a speaker. As with speech recognition, humans are the most accurate language identification systems in the world today. Within seconds of hearing speech, people are able to determine whether it is a language they know. If it is a language with which they are not familiar, they often can make subjective judgments as to its similarity to a language they know. Any utterance is nothing but a speech or audio signal. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing.

There are different aspects incorporated in speech which can be employed to represent the characteristics of a language. The raw speech signal is complex and may not be suitable for feeding as input to the language identification system, hence the need for a good front-end arises. The task of this front-end is to extract all relevant acoustic information in a compact form. In other words, the pre-processing should remove all non-relevant information such as background noise, and encode the remaining (relevant) information in a compact set of features that can be given as input to the classifier.

The major task is to identify what features have to be extracted in order to discriminate between languages. Feature is a broad term with respect to speech signals. They could be acoustic features, prosodic features or phonotactic features. Prosody is the rhythm, stress, and intonation of speech. The prosodic of oral languages involve variation in syllable length, loudness, pitch, and the formant frequencies of speech sounds. This includes phoneme length and pitch contour. These prosodic units are the actual phonetic "spurts", or chunks of speech. Phonotactic: are rules that govern permissible sequence of phonemes in speech signals. Phonotactic defines permissible syllable structure, consonant clusters, and vowel sequences by means of phono tactical constraints.

The acoustic features are the low level features from which the prosodic and phonotactic features are derived. The acoustic features deal with modelling those parameters which are obtained from digital signal processing techniques. The power spectrum of a signal is indicative of acoustic information in speech. We make use of the Cepstral analysis of the power spectrum of the speech signal. A Cepstrum is the result of taking the Inverse Fourier transform of the logarithm of the spectrum of a signal. This data is used to model the language feature space.

II. Problem Statement:

- The Language Identification Systems should identify the language irrespective of gender, accents and pronunciations.
- An Acoustic model is used to extract suitable features from speech samples across different languages and different speakers in each language.
- The language set considered in this project involves 4 – Indian Languages namely Assami, Bengali, Gujarati and Hindi.

III. Models Implemented for LID system

- GMM based Bayes Classifier model
- Fully Connected Neural Network model
- Convolutional Neural Network model

IV. System Design

DNN plays a major role in speech enhancement by creating a model with a large amount of training data and the performance of the enhanced speech is evaluated using certain performance metrics. DNN's inherently fuses the process of feature extraction with classification and enables the decision making.

- It can read directly from text, sound, or images and can achieve incredible accuracy, so gives better results than traditional machine learning techniques
- The possibility of using the voice identification with no database is one of the major feature of DNN.

Feature Extraction:

- It is a process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal.
- Aims to reduce the number of features in a dataset by creating new features from the existing ones.
- It reduces the magnitude of the speech signal devoid of causing any damage to the power of speech signal.

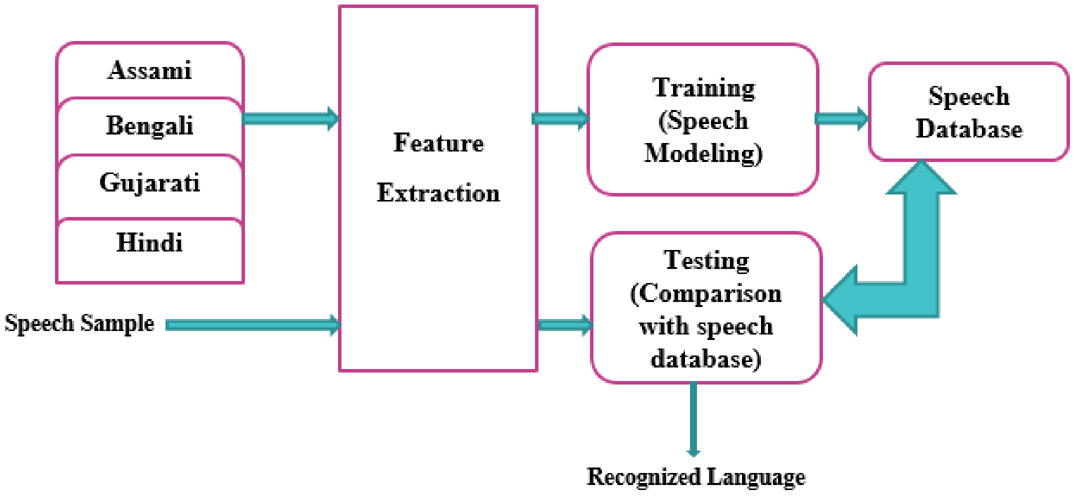


Fig 4.1 Working of LID System

MFCC - Mel-Frequency Cepstral Coefficient:

- MFCC takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale,
- Mel Scale Relate the perceived frequency of a tone to the actual measured frequency.
- It can compute formants that are in the low frequency range and describe the vocal tract resonances.

MFCC Function

- Separates the audio into short window/frame and calculates the MFCC for each frame (Feature Vector)
- It provides the shape of mfccs, how many mfccs are calculated on how many frames.
- The first value represents the number of mfccs calculated and another value represents a number of frames available.

Role of MFCC?

- Speech is represented as sequence of Cepstral vectors (MFCC Vectors)
- These MFCC feature vectors are given to pattern classifier for different speech processing task

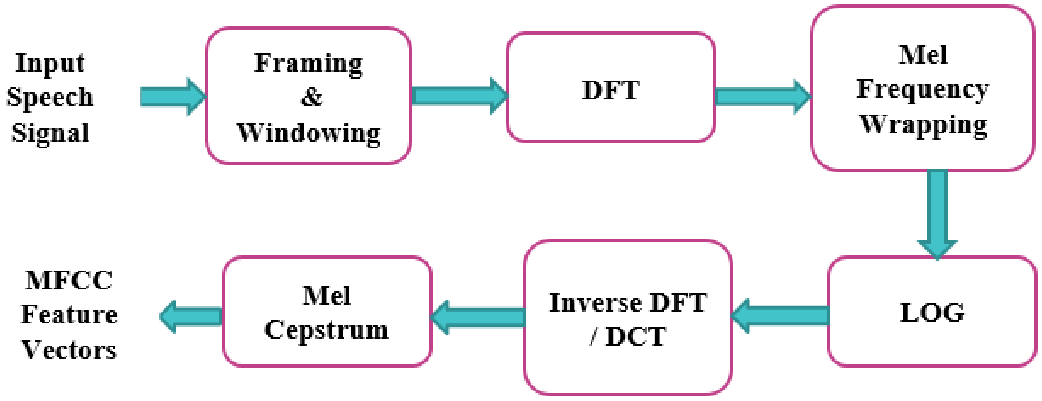


Fig 4.2 MFCC Feature Extraction

Spectrogram:

- It is a way of representing the signal strength/loudness over time at various frequency band
- Spectrogram is a visual representation of a spectrum of a sound changing through time.

How to calculate Spectrogram?

- Divide the signal into equal length segments
- Window each segment and compute its spectrum to get the short time fourier transform
- Display segment by segment the power of each spectrum in decibels.

Mel Filter Banks:

- Triangular filter banks that helps to capture the energy at each critical frequency band and roughly approximates the spectrum shape.
- Helps to smooth the harmonic structure of a spectrum.

Cepstrum:

It is a spectrum of spectrum or Mel spectrum.

V. GMM BASED BAYES CLASSIFIER FOR LID SYSTEM

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

Advantage of GMM

- It allows the model to learn the subpopulations automatically.
- GMM tends to group the datapoints belonging to a single distribution together.

Design steps:

- Training set of speaker includes ‘N’ number of speech signals.
- Each speech signal is represented as set of ‘D-dimensional’ MFCC feature vector.
- GMM is obtained by clustering all these MFCC feature vectors (into ‘Q’ number of clusters)

Training Objective:

- Estimate the parameters of the GMM using maximum likelihood method.

VI. FCNN FOR LID SYSTEM

- Fully Connected Neural Networks are proven as a good classifier
- The major advantage of FCNN is that they are "structure agnostic. (No special assumptions need to be made about the input)

Design steps:

- Training set of speaker includes 'N' number of speech signals.
- Each speech signal is represented as set of 'D-dimensional' MFCC feature vector.
- FCN will train to identify the language by using MFCC feature vector of the training set

Training Objective:

- To Identify the Audio MFCC feature vector to Classify the Languages.

VII. CNN FOR LID SYSTEM

- CNN is more accurate in natural language processing and is more efficient to achieve training results.
- Convolutional Neural Networks learn patterns that are translation invariant and have spatial hierarchies.
- The sound files are converted into spectrograms, then the CNN Classifier model will produce predictions about the class to which the sound belongs.

Design steps:

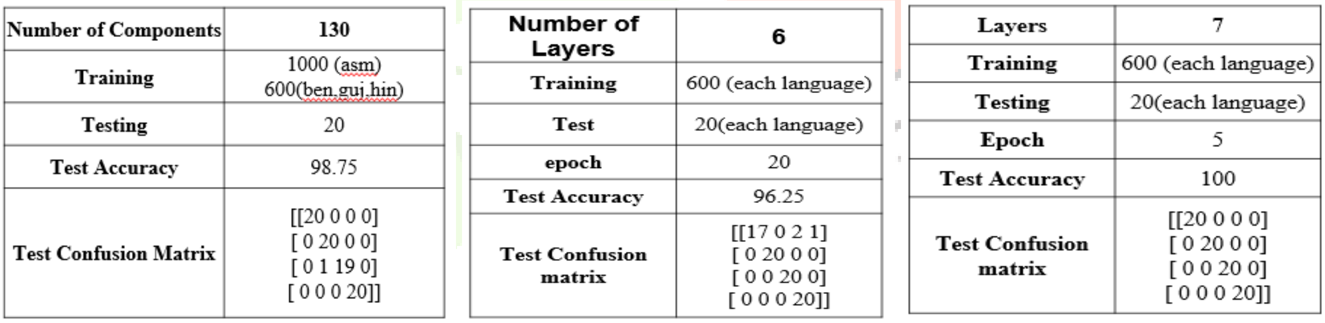
- Training set of speaker includes 'N' number of speech signals.
- Each speech signal is represented as set of 'D-dimensional' Spectrum.
- CNN will train to identify the language by using the Spectrum of the training set

Training Objective:

- To Identify the Audio Spectrum to Classify the Languages.

VIII. RESULTS AND DISCUSSION

The datasets for all our experiments are randomly taken from different parts of Web like podcasts and online audio books. The datasets are divided into two parts: Training Data and Testing Data. MFCC feature extraction is done to training the machine for different languages. The system is trained over a large corpus of data and a small subset is used for testing to achieve better accuracy. The experiments are conducted to analyze the response of the proposed LID against the considered 4 - different Indian languages (Assami, Bengali, Gujarati and Hindi). The results are as depicted in following table.



Number of Components	130
Training	1000 (asm) 600(ben.guj.hin)
Testing	20
Test Accuracy	98.75
Test Confusion Matrix	[[20 0 0 0] [0 20 0 0] [0 1 19 0] [0 0 0 20]]

Number of Layers	6
Training	600 (each language)
Test	20(each language)
epoch	20
Test Accuracy	96.25
Test Confusion matrix	[[17 0 2 1] [0 20 0 0] [0 0 20 0] [0 0 0 20]]

Layers	7
Training	600 (each language)
Testing	20(each language)
Epoch	5
Test Accuracy	100
Test Confusion matrix	[[20 0 0 0] [0 20 0 0] [0 0 20 0] [0 0 0 20]]

Table 8.1 Results of GMM, FCNN & CNN Models

Further, experiments are conducted to demonstrate the system accuracy for all 4 – Indian languages. Around 1600 samples are fed to the system and the LID demonstrated around 100% classification accuracy using CNN Model. The graph of classification accuracy of the system is shown in Figure 8.1 and it is evident that the systems perform well as it comes across more evidences against each language.

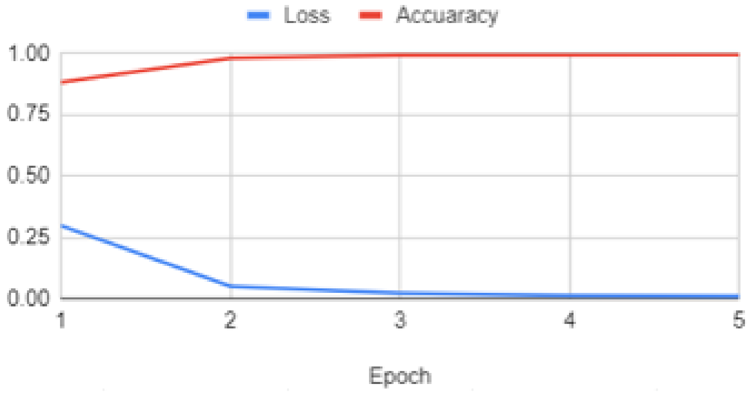


Fig 8.1 LID Performance for CNN

IX. Acknowledgment

We are thankful to ATMECE, Mysore, CSE Department for having provided us with an opportunity to implement this project. Also we would like to thank our principal Dr. L Basavaraj and HOD, Department of CSE for this opportunity.

References

[1] "Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal" Lukas Mateju,Pet Cerva, Jindrich Zdansky, Radek Safarik Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic {lukas.mateju, petr.cerva, jindrich.zdansky.

[2] https://towardsdatascience.com/deep-neural-network-language-identification-ae1c158f6a7d

[3] https://paperswithcode.com/task/spoken-language-identification

[4] https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning

[5] https://github.com/ibro45/Language-Identification-Speech

[6] Spoken language identification based on the enhanced self-adjusting extreme learning machine approach Musatafa Abbas Abbood Albadr1*, Sabrina Tiun1, Fahad Taha AL-Dhief2, Mahmoud A. M. Sammour3