IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 11, Issue 7, July 2021, ||Series -II|| PP 24-30

Malayalam Speech to Text Conversion Using Deep Learning

Arun HP[1], Jithin Kunjumon[2], Sambhunath R[3], Ancy S Ansalem[4]
B. Tech.Student[1,2,3], Assistant Professor[4]
Department of Electronics and Communication Engineering Mar Baselios College of Engineering and Technology, Thiruvananthapuram, Kerala, India
Received 20 July 2021; Accepted 5 August 2021

ABSTRACT— In the current scenario, speech recognition for several languages is becoming more popular. Recognizing speech is a very difficult task in the Malayalam language. This project aims to establish a Formal Malayalam Speech to Text converter for the language of Malayalam. The system considers only isolated words with constrained vocabulary. The word which is spoken by the speaker is given as the input to the system is presented in the display as the output. We are using deep learning and feature extraction techniques for this project. The proposed system is taking around 5-10 isolated words for tutoring the machine. Since the system is depending on the speaker voice, at the beginning the words are stored in .wav (waveform audio) file for training procedure. Several samples are stored and trained for each word. The input audio word will be collated with these stored words. Pre-processing process includes the transformation of speech signal into digitized format. This digital signal is passed to the first order filters for the smoothening signals, which would help in the rise of signal's energy at a higher frequency. MFCC is the systematic technique for feature extraction. Mel-frequency cepstral coefficients are obtained, after the completion of this phase. MFCC examines the frequencies with human perception sensitivity. Following the pre-processing, syllabification, and feature extraction procedure, HMM is used to identify the speech and training. The speech recognition system based on ANN was implemented using LSTM which is a common form of neural network.
KEYWORDS: Constrained vocabulary, ANN, LSTM, HMM, MFCC, Extraction techniques, Deep learning, Pre-processing, Syllabification.

I. INTRODUCTION

Speech is considered as the one of I. the common modes of communication among human beings in the modern civilized societies and it is the most natural and efficient way of exchanging information. Usually, users get interacted with computer via keyboard and mouse. If there is a large quantity of data to be recorded, it takes more time to get processed. So, if a system can understands the human language, then it will be the best form of interaction between a human and a computer. A speech to text conversion system takes audio signal as the initial data, identifies it, and transform it into text. It encourages numerous applications which includes a helping hand for illiterate individuals, support of telephonic directories, supervision gadgets for updating the health status in hospitals, in industrial banking sector etc. Malayalam is one among the 22 languages spoken in India with about 40 million speakers. Malayalam is one of the Dravidian languages and is the official language of Kerala and the Union territory Lakshadweep. There are 37 consonants and 16 vowels in the language. It is a syllable-based language and written with syllabic alphabet in which all consonants have an inherent vowel /a/. There are different forms in Malayalam even though the literary dialect throughout Kerala is almost uniform. Many speech to text recognition works has been taken place in many of the Indian languages and foreign languages. Nevertheless, very less work has been done in Malayalam.

In this paper, we are presenting a speech to text conversion scheme using deep learning techniques for the Malayalam language. Here, our system is studying 5-10 secluded words for the training. At first, the phrases are collected and trained. For each word, record a number of models and store it. The phrase pronounced will be compared with these stored words. In most of the speech recognition systems, the acoustic modelling components of the recognizer are almost exclusively based on HMM. Hidden Markov model provides an elegant statistical framework for modelling speech patterns using a Markov process that can be represented as a state machine. The probability distribution associated with each state in an HMM, models the variability which occurs in speech across speakers or even different speech contexts. Speech Pre-processing, Feature extraction and speech classification are the important stages in this system. Some of the feature extraction techniques used in Speech to text conversion systems are Linear Predictive Coding (LPC), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Principal Component Analysis (PCA), MFCC, Kernel based feature extraction, Wavelet Transform and spectral subtraction. The most generally used technique is MFCC. For the process of recognizing speech, Hidden Markov Model, Artificial Neural Networks (ANN) are various techniques used. After pre-processing, syllabification and feature extraction, HMM is used to realize the speech

and training. The system is being trained by means of task grammar, acoustic models using HTK toolkit. The speech recognition system based on ANN was implemented using LSTM which is a popular form of neural network. The software used for testing and training is PyCharm which is an integrated development environment used in Python programming and TensorFlow is used as an open-source software library for machine learning which help to focus on training and inference of deep neural networks.

II. RELATED WORKS

Deep Learning is nowadays widely used in automatic speech recognition (ASR) systems. At the earlier stages of 1950's lots of diverse techniques and methodologies were recommended considering variant emerging issues and applications. The variability of ASR technologies mainly depends on two categories namely: Acoustic Feature Extraction process, Recognizer or Classifier. On taking the feature extraction procedure, most functional and more accurate is Mel frequency Cepstral Coefficients (MFCCs). Other includes Linear Predictive Coding (LPC), Linear Pre- dictive Cepstral Coefficients (LPCCS), Perceptual Linear Pre-dictive features (PLP), Discrete Wavelet Transform (DWT) etc. Also, there are many ASR classifiers have been proposed namely: Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Discrete-Time Warping (DTW), Viterbi algorithm; Support Vector Machine (SVM) and Vector Quantization (VQ) etc. Among those HMM classifier is considered as an excellent classifier for ASR process.

Suma Swamy et al. introduced an efficient speech recognition system which was tested with Mel Frequency Cestrum Coefficients (MFCC), Vector Quantization (VQ), HMM which identify the speech by 98% accuracy in 2013. The database consists of five words spoken by 4 speakers at ten times. In 2017, Preena Johnson, Jishna K, Soumya established a Malayalam word identification for speech recognition system. The proposed system is trained for five words. Every word has 5 samples to be taped. The results for the phrase are examined 25 times and accuracy percentage is determined. The system is granting an accuracy of about 75% when demonstrated using GMM. In 2012, Ms. Vimala. C and Dr. Radha has planned a speaker independent isolated language recognition system for Tamil language. The dataset used is 10 Tamil spoken digits (0-9) and 5 spoken names from 30 distinct speakers. Those data are sampled at a rate of 16 Khz. Their approach toward the Tamil language recognition system has gained an accuracy of 88% in 2500 words and FWCMN-MTYWGFCC-FF strengthened the WRR up to 99.06% for the HMM Techniques.

III. METHODOLOGY

At first the Malayalam dataset is created by recording and storing the words in .wav file. For recording, in Python, we are using Pyaudio, which has inbuilt audio related functions. For each word, a number of samples are to be recorded by different speakers as separate files. The parameters are to be specified in the audio stream includes frame size, format, channels, rate etc.

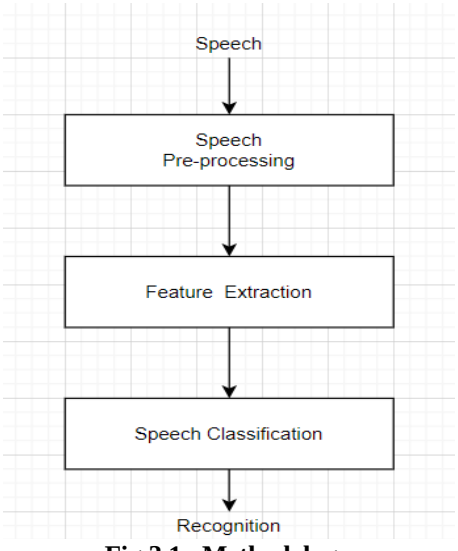

Fig 3.1 Methodology

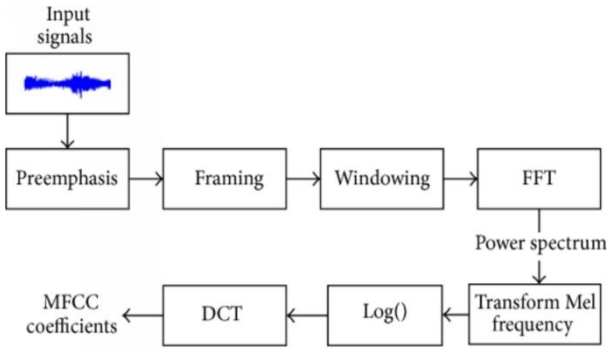

Fig 3.2 MFCC Block Diagram

Feature Extraction

For recognition of speech, the signals have to be represented with some specific features. MFCC is the well-known popular method of feature extraction. To capture the phonetically important characteristics of speech, signal is expressed in Mel-Frequency Scale. This scale has a linearly frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. MFCCs are less susceptible to the physical conditions of the speakers' vocal cord, compared to the speech wave forms. The block diagram of the feature extraction process is shown in figure 3.2. After the signal being pre-processed, speech waveforms fed into frame blocking and windowing process. Then the time domain signal is converted into frequency domain by applying Fast Fourier transform (FFT) on it. Then the spectrum is fed into Mel frequency wrapping. This involves two steps: -Mel-scale and filter banks. Here, for each tone of the signal, a subjective pitch is measured on the 'Mel'. For a given frequency f, measured in Hz, mels are calculated by

$$\text{mel}(f) = 2595 \log 10(1 + f/700) \quad \ldots\ldots. \text{(i)}$$

Mel Spectrum coefficients has to be converted to the time domain by applying Discrete cosine Transform. Discrete cosine transform makes a transformation from the frequency domain into a time-like domain termed as quefrency domain. The features attained are like a cepstrum, thus it is referred as the Mel-scale cepstral coefficients. Mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low que-frequency peaks. MFCC is calculated as:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)); \ldots \quad \text{(ii)}$$

$$n = 0,1,2,\ldots.-1$$

where c(n) are the cepstral coefficients, and C is the number of MFCCs. Generally, MFCC systems use only 8–13 cepstral coefficients. For each speech frame of about 25ms with overlap, a set of Mel-frequency cepstrum coefficients are computed. These set of coefficients are called an acoustic vector. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.

```
E:\speechmfcc\venv\Scripts\python.exe E:/speechmfcc/mfcc.py
Sample rate: 16000Hz
Audio duration: 4.9s
Framed audio shape: (327, 2048)
First frame:
[-6. -1.  3. ... -7. -9. -7.]
Last frame:
[-18. -19.  -8. ...   1.  -2. -13.]
(327, 1025)
Minimum frequency: 0
Maximum frequency: 8000.0
MEL min: 0.0
MEL max: 2840.023046708319
(10, 327)
[ 1.02848995e+02  2.23805833e+01  1.90258223e+00  2.29259815e+00
  1.25864600e+00 -1.86399369e+00 -8.07545853e-01  3.18193782e-02
  5.40562317e-01 -7.80046078e-01 -1.12506671e-13  7.80046078e-01
 -5.40562317e-01 -3.18193782e-02  8.07545853e-01  1.86399369e+00
 -1.25864600e+00 -2.29259815e+00 -1.90258223e+00 -2.23805833e+01
 -1.45450444e+02 -2.23805833e+01 -1.90258223e+00 -2.29259815e+00
 -1.25864600e+00  1.86399369e+00  8.07545853e-01 -3.18193782e-02
 -5.40562317e-01  7.80046078e-01  9.39772011e-15 -7.80046078e-01
  5.40562317e-01  3.18193782e-02 -8.07545853e-01 -1.86399369e+00
  1.25864600e+00  2.29259815e+00  1.90258223e+00  2.23805833e+01]

Process finished with exit code 0
```

Fig 3.3 Cepstral Coefficients

Speech Classification

Classification is the technique of matching input with the model created. In this work, HMM and LSTM are used as a classification technique. Hidden Markov model is used as a classifier to compare the extracted features from MFCC with stored templates. An unknown speech wave form is converted by a frontend signal processor into a sequence of acoustic vectors, O = o1 o2, o3, o4 …o t. The utterance consists of sequence of words W = w1, w2, w3 …. w n. In ASR, it is required to determine the most probable word sequence, S, given the observed acoustic signal O. Applying Bayes' rule,

$$S = \arg\max P(W / O) \quad \text{……. (iii)}$$

Hence a speech recognizer should have two components: P (W), the prior probability, is computed by language model, while P(O/W), the observation likelihood, is computed by the acoustic model. In this work, the acoustic modelling is computed by HMM. Since HMM is a statistical model in which it is assumed to be in a Markov process with unknown parameters, the challenge is to find all the appropriate hidden parameters from the observable states. Hence it can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a Hidden Markov model, the state is not directly visible, while the variables influenced by the states are visible. Each state has a probability distribution over the output. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. Thus, HMM model can be defined as:

$$\lambda = (Q, O, A, B, \Pi) \quad \text{……. (iv)}$$

where, Q is $\{q_i\}$ (all possible states), O is $\{v_i\}$ (all possible observations), A is $\{a_{ij}\}$ where

$$a_{ij} = P(X_{t+1} = q_j \mid X_t = q_i) \text{ (Transition probabilities),}$$

B is $\{b_i\}$ where $b_i(k) = P(O_t = vk \mid X_t = q_{it})$ (Observation probabilities of observation k at state i), $\Pi = \{\pi_i\}$ where $\pi_i = P(X_0 = q_i)$ (Initial state Probabilities), and O_t Denote the observation at time t.

IV. TESTING AND TRAINING DETAILS

Training is done by Long short-term memory network. In training phase knowledge models are created for the phonetic units. The database is divided into three equal parts and for each experiment, 2/3 of the data is selected for training and the remaining 1/3 is selected for testing. From the test results word accuracy rate for each set is calculated. Using the trained model, the system has also tested with speech from unknown speakers. Using the TensorFlow backend, our team developed ASR model using LSTM neural network The model training done by NVIDIA GeForce 940MX, 4GB Dedicated Graphics, CUDA Enabled GPU. The model was trained with 360 audio samples. Learning rate is set to 0.0001 and 3000 steps of training iterations. Training a neural network is a weight adjustment in ANN to create a model for prediction. The assessment of effectiveness of deep learning can be done in many ways. The popular method which can measure the accuracy of the deep learning work is to measure accuracy and F1 score measurement. The equation (v) is a measure of accuracy and correctness.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{……... (v)}$$

After predicting the binary classification results, the possible results are positive and negative. Then the results would be as follows: Predictions are true positive (TP), true negative (TN), false negative (FN), and false positive (FP). F1 measurements can be calculated from below equation (vi).

$$F1 = 2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right) \quad \text{……(vi)}$$

And the $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{recall} = \frac{TP}{TP+FN}$

V. EXPERIMENTS AND RESULTS\

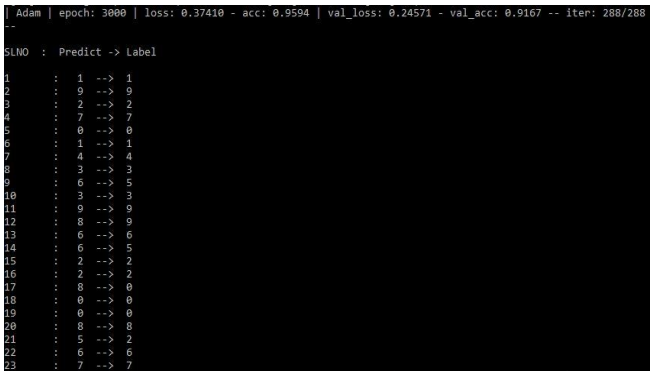

Fig 5.1 Prediction Result

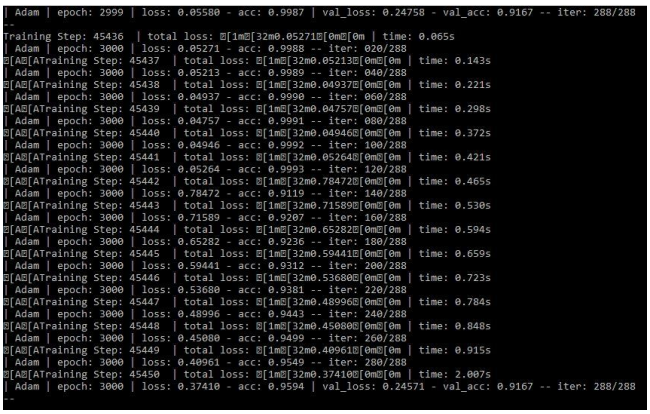

Fig 5.2 Training

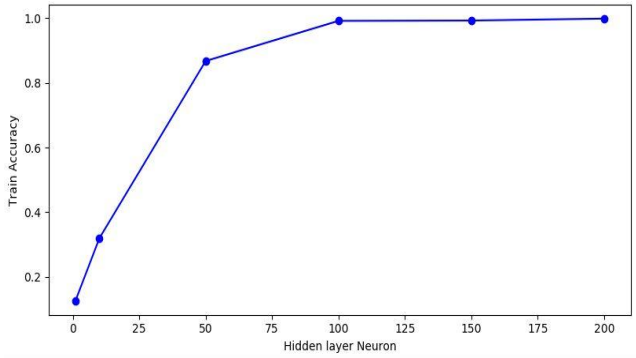

Fig 5.3 Training Accuracy vs No of neurons

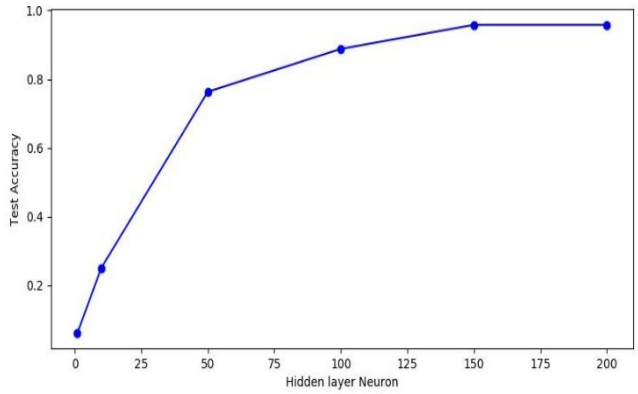

Fig 5.4 Testing Accuracy vs No of neurons

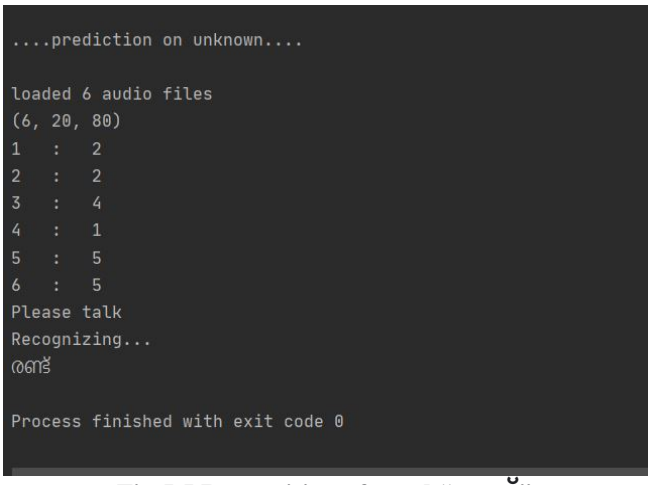

Fig 5.5 Recognition of word "രണ്ട്"

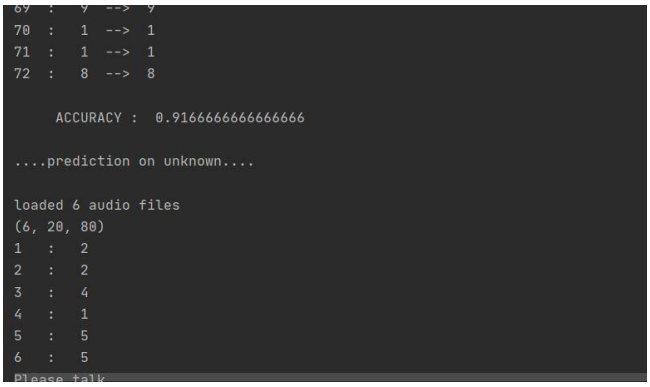

Fig 5.6 System Performance

VI. CONCLUSION

An ASR system is modelled as an initiative towards an advanced speech to text conversion system for Malayalam. Our system is giving an accuracy of about 91% when modelled using HMM classification and LSTM training. On concluding that HMM based MFCC feature is more suitable for speech recognition requirements and produces more good results than other models. We would like to enlarge this project to a speaker independent system which also deals with a large vocabulary system with continuous and connected words. It also helps the illiterate people to write Malayalam words and we promote the use of our native language in our daily life.

REFERENCES

[1]. C. Kurian and K. Balakrishnan, "Speech recognition of Malayalam numbers", Nature & Biologically Inspired Computing (NaBIC), pp. 1475-1479, 2009.

[2]. Neha Chauhan, Mahesh Chandra, "Speaker recognition and verification using Artificial neural network", IEEE Conference The international Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) pp.1173-1176, 22-24 March, 2017.

[3]. Neha Chauhan, Tsuyoshi Isshiki, Dongju Li," Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database", IEEE, ICCCS Singapore, Feb,2019.

[4]. Vimala. C, V. Radha, "Efficient Acoustic Front-End Processing for Tamil Speech Recognition using Modified GFCC Features", International Journal Image, Graphics and Signal Processing, DOI: 10.5815/ijigsp.2016.07.03, 2016, pp.22-31

[5]. Dr. C. Kurian, "Speech database and text corpora for Malayalam language automatic speech recognition technology", Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques(O-COCOSDA),2016. doi:10.1109/icsda.2016.7918975

[6]. Virendra Chauhan, Shobhana Dwivedi, Pooja Karale, Prof. S.M. Potdar," Speech to text converter using Gaussian Mixture Model (GMM)", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 02 | Feb-2016

[7]. "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)" Su Myat Mon, HlaMyoTun, International Journal of Scientific & Technology Research ISSN 2277-8616, vol4, issue 06, JUNE 2015
[8]. Banerjee, Adrish, Dubey, Akash Menon, Abhishek, Nanda, Shubham, Nandi, Gora Chand, "Speaker Recognition using Deep Belief Networks", eprint arXiv:1805.08865, May 2018.
[9]. Parashar Dhakal, Praveen Damacharla, Ahmad Y. Javaid and Vijay Devabhaktuni," A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface", Machine. Learning and Knowledge. Extraction, 2019, 1(1),504-520.

Arun HP, et. al. "Malayalam Speech to Text Conversion Using Deep Learning." *IOSR Journal of Engineering (IOSRJEN)*, 11(07), 2021, pp. 24-30.