COMPARATIVE STUDY ON DIFFERENT ARCHITECTURES USED FOR THE AUTOMATIC SPEECH RECOGNITION OF MALAYALAM- A LOW RESOURCE LANGUAGE

Manju G [1], Anil Kumar K S [2]

[1]*Department of Future Studies, University of Kerala, India*
E-mail:manjualoshious@gmail.com
[2]*Department of Future Studies, University of Kerala, India*
E-mail:ksanilksitm@gmail.com

Abstract

Speech is the most natural form of human communication and is one of the most information-laid signals. Sound waves have high and multi-layered temporal-spectral differences. In ASR acoustic features are mapped to the basic units of sound called phonemes, and then reconstructed into words and sentences of a particular language. Malayalam belongs to the Dravidian family of languages. A Speech recognition programs have a broad range of applications. Malayalam is a low resource south Indian language because of lack of availability of enough speech-to-text corpora. It also faces challenges like code switching, different accents in different regions, non-native speakers etc. Development of Malayalam speech recognition system is at its beginning stage. Only a few researches were done in this field. Speech features can be extracted and modeled using different methods. The purpose of feature extraction is the reduction of the high dimensionality of audio signal without losing the quality of speech data. The paper reviews the feature extraction and modeling techniques used in Malayalam Speech Recognition systems.

Keywords: Malayalam Speech Recognition, feature extraction, MFCC, PLP, SVM.

1. INTRODUCTION

The most natural form of human communication is speech. Speech is a sequence of elementary acoustic sounds known as phonemes and it is one of the most information-laid signals. Sound is spectral mix of harmonic components, noise, and silent. Speech signal have rich and multi-layered temporal-spectral changes and includes time and frequency modulation of information as formants and pitch intonation [1]. A **formant** is a concentration of acoustic energy around a particular frequency in the speech signal. **Intonation** is the changes of spoken pitch and is not used to differentiate words; alternatively, it is used for identifying the attitudes and sentiments of the speaker, distinguishing statements and questions, and differentiating different categories of question, focusing the attention etc. [1]

Acoustic phonetic symbols are the basic speech units from which other speech units such as syllables and words are formed. The information contained in audio signals carries many features. The rhythms of speech called prosody helps to convey information as the boundaries between fragments of speech, connect sub-phrases and elucidate the purpose and remove ambiguities such as whether a spoken sentence is a statement or question [2]. Gender is conveyed by the pitch, the size and physical characteristics of the vocal tract. Female voice has higher resonance frequencies and a higher pitch than male voice because of shorter vocal cords. The size and the elasticity of the vocal cords and vocal tract and the pitch are the major components that covey the age of a person. The accent of a speaker is conveyed through changes in the articulation dictionary in the form of substitution, deletion or insertion of phoneme units in the "standard" transcription of words and systematic changes in speech formants, pitch intonation, duration, emphasis and stress [2]. Speaker recognition is characterized by the physical characteristics of an individual's larynx, vocal tract, pitch intonations and stylistics. Emotion and health, conveyed by changes in vibrations of vocal fold, resonance frequency, duration and stress [2].

2. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition (ASR) is the process of converting acoustic signals into written text. In ASR acoustic features are mapped to the building units of voice called phonemes, and words and sentences of a particular language are constructed from the phonemes. Speech recognition system can be classified by what type of utterances they can recognize as [3],

a. Isolated Word -Isolated word recognizer requires each word to have silence on both side of it.
b. Connected Word -Connected word systems are similar to isolated words but allow many words to be speak together with minimum pause between the words.
c. Continuous speech - Continuous speech recognizers allows user to speak almost naturally where the context is already determined. Building ASR system for continuous speech recognition is more difficult than isolated and connected word recognition.
d. Spontaneous speech -It is most natural form of speech. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech.

The ASR systems that use whole-word matching compares the incoming digital-audio signal with a word template that is already recorded. This technique takes very less time for processing than sub-word matching technique, but the user has to construct a pronunciation dictionary which consists of every word that that is to be recognized. These type of ASR systems require large volume of storage space and is practical only in case of context sensitive vocabulary. But in sub-word matching, the system looks for sub-words like phonemes to extract features and then performs further pattern recognition on these

features. This technique takes more computation time than whole-word matching, but it needs only very less storage [3].

There is a rapid development in Automatic Speech Recognition (ASR) in the recent years. The majority of these achievements are constrained to languages having high volume annotated speech corpus. But low resource South Indian languages like Tamil, Malayalam etc., always have challenges in adopting the same methodology used for languages having massive dataset for training and testing the ASR model. Additionally, the low resource languages face challenges like multilingualism, less fluent native speakers for data collection, and too many accents etc. [4].

3. MALAYALAM – A LOW RESOURCE LANGUAGE

Malayalam is a language spoken by the people of Kerala, a state of Southern India. Malayalam is a morphologically complex language with 16 vowels and 36 consonants. It has seven nominal case forms, two nominal number forms and three gender forms. These forms are used as suffixes to the nouns for nominal inflection. Tense, mood, voice and aspect causes verb inflection [5]. Malayalam also has a canonical word order of Subject- Object- Verb (SOV) agreement.

Like any other regional languages in India, Malayalam is a low resource south Indian language in terms of availability of speech-to-text corpora. Malayalam language has scarcity of linguistic knowledge with technological know-how, annotated speech corpus, and various language models. Malayalam also has many accents in different areas of the state, language switching, and many non-native speakers [6].

4. MALAYALAM ASR -RELATED WORKS

Only a few research works were done in the area of automatic Malayalam speech recognition.

Sharika - an Automated Speech Recognition (ASR) system is an initial initiative in Malayalam speech recognition. Hidden Markov Models (HMM) is used for recognition process by employing MFCC. It was developed as simple 8 command system which is then enhanced to 50 commands. This project uses the Sphinx engines developed by Carnegie Mellon University (CMU) [7].

The first work on Speaker Independent Malayalam Isolated Digit Recognition is based on Perceptual Linear Predictive Cepstral Coefficient (PLP) feature extraction and Hidden Markov Model (HMM) for ASR modeling. The system obtained a recognition accuracy is 99.5% in recognizing 10 digits spoken by 21 speakers [8].

Another work reported compares and evaluates the performance of context dependent tied (CD tied) model, context dependent (CD) model and Context independent (CI) models for the continuous speech recognition of Malayalam language [9]. In this work the dataset for the speech recognition system has utterances from 11 females and 10 males. The evaluation results show that CD tied models performs better than CI models over 21%. And the maximum Sentence Recognition Accuracy is 81.5%. The system is designed using Phoneme-based Hidden Markov Models (HMM) and MFCC features.

A related work handles speaker independent connected digit speech recognition. The ASR system uses Perceptual Linear Predictive Cepstral coefficient for extracting speech features. The ASR system is modeled using continuous density Hidden Markov Model. The training data base uses the utterance of 21 speakers from the age group of 20 to 40 years and the sound is recorded in the normal office environment. Each speaker read 20 set of continuous digits [10] and the recognition accuracy obtained by the system is of 99.5%.

Another work is the development of a speaker independent Automatic Speech Recognition System for isolated Malayalam vocabulary [11]. In this a hybrid system consisting of wavelet packet decomposition and artificial neural networks is used for recognizing speaker independent isolated spoken words in Malayalam and the overall recognition accuracy obtained is 87.5%

A recent work is an Acoustic Model for Isolated Malayalam Speech Recognition with different Gaussian Mixtures [12]. It focused on the development of an acoustic model for medium vocabulary, context independent, isolated Malayalam Speech Recognizer. the ASR system used Hidden Markov Model (HMM) for acoustic modelling using MFCC feature extraction method. A speech dataset used for the speech recognizer consists of 100 words spoken by 6 speakers including both male and female. The ASR system shows a accuracy of 90.91% for Single Gaussian and the Word Error Rate (WER) obtained is 11.9%. When the experiment is conducted for multivariate Gaussians for Gaussian Mixture five it attained a word accuracy of 95.24% with WER of 4.76%

Another work is Malayalam Speech to Text Conversion modelled using Deep Learning. The ASR system used HMM for the classification and LSTM for the training of isolated words with constrained vocabulary [13].

Another work developed an Automatic Speech Recognition system on Malayalam speech data using spectrogram images. It used Convolutional Neural Network (CNN) for acoustic modeling and obtained an accuracy of 93.33%. The Convolutional Neural Network is built with a set of Convolution and Fully Connected layers. The CNN used SoftMax layer for classification of speech data. 4000 tokens were used by ASR system and it showed that spectrogram image-based approaches have favorable results in voice recognition [14].

Another work on developing a conventional and syllable-based ASR systems for Malayalam used DNN for speech modeling. [15]. The analysis shows that Kaldi performed well for phoneme-level DNN acoustic modeling using MFCC features, resulting a lower WER of 2.86% than the syllable-based approach.

Another work designed a formal Malayalam Speech to Text converter for isolated words with limited vocabulary. The system is designed to recognize around 5-10 isolated words by using deep learning and MFCC feature extraction technique [16].

A related research work developed a Convolutional Neural Network-Based Automatic Speech Emotion Recognition System for Malayalam language. The Emotion Recognition system used CNN and deep learning techniques for the modelling of the language. This paper addresses the challenges in sentimental analysis of low resource language by designing Long Convolutional Neural Networks (CNN) to recognize sentiments in audio dataset of Malayalam. The Mel Frequency Cepstral Coefficient (MFCC) techniques is used to extract features from the voice data [17].

Malayalam is a morphologically complex language. A recent research work developed an automatic speech recognition system using sub word tokens for language modeling. The speech recognition system developed for Malayalam ASR used a DNN-HMM (Deep Neural Network–Hidden Markov Model) based automatic speech recognition with MFCC approach [18].

Another work developed an ASR system using support vector machine (SVM) for vowels in the Malayalam language. The experiment was conducted using speech samples of children in the age group of five to ten. The work reported an accuracy of 88% using the SVM with Quadratic. the system showed a recognition accuracy of 89.5% for Cubic, 77.5% for Fine Gaussian, 91.5% for Medium Gaussian and 82% for Coarse Gaussian kernel functions. MFCC is used for extracting features [19].

Table 1: ASR systems for Malayalam Language

Author	**Feature Extraction Methodology**	**Classifier**	**Accuracy**	**Dataset feature**
[7]	MFCC	HMM	-	Isolated Word, Limited Vocabulary
[8]	PLP Cepstral coefficient	Continuous density HMM	99.5%	Isolated Malayalam Digit, Limited Vocabulary
[9]	MFCC	HMM, CD tied	81.5%	Continuous Digit, Limited Vocabulary
[10]	PLP Cepstral coefficient	HMM	99.5%	Connected digit
[11]	Wavelet Packet Decomposition	ANN	87.5%	Isolated Malayalam Word
[12]	MFCC	HMM & GMM	Single Gaussian - 90.91% Word Error Rate - 11.9% Multivariate Gaussians -95.24%	Medium vocabulary, speaker dependent isolated Malayalam speech corpus of 100 words
[13]	MFCC	Deep Learning model using HMM classification and LSTM	-	context independent, isolated Malayalam Words
[14]	Spectrogram Image	CNN	93.33 %	4000 tokens
[15]	MFCC	DNN	Higher WER than Isolated Word Recognition	Syllable based. Limited vocabulary
[16]	MFCC	HMM classification and LSTM	91%	isolated words with constrained vocabulary of 5-10 isolated words.
[17]	MFCC	CNN	71%	Limited Vocabulary for Sentimental Analysis
[18]	MFCC	DNN-HMM	10.6% WER	Limited Vocabulary

[19]	MFCC	SVM	SVM - Quadratic (88%), Cubic (89.5%) Fine Gaussian (77.5%), Medium Gaussian (91.5%), and Coarse Gaussian (82%)	Vowels in Malayalam, Smaller data set

5. SPEECH SIGNAL PROCESSING –AN OVERVIEWS

Sound is produced from objects that vibrate and exerts pressure changes in a sound-transmitting channel like air. The sound is produced by the vibrations of the air in the glottis. The smallest frequency that is produced by a vibrating body is called the fundamental frequency and the vibration occurs in harmonic. All other frequencies present in the sound wave are multiples of the fundamental frequency. Based on the position of the tongue and other articulators the frequencies are get amplified and reduced by the vocal cavity. The fundamental frequency provides information about the pitch and other frequencies provide information about the phones. When a sound waves propagating outward from a vibrating body when reaches the eardrum of a listener, the eardrum vibrates and the process of hearing is started off. Along with the message conveyed, speech data contain different speaker specific information. [20].

The speech processing begins by converting speech signal to digital form with a sampling frequency. When a computer records digital audio, it measures the sound pressure level multiple times per second. These measurements are often called samples. The Nyquist-Shannon theorem for sampling states that the minimum sampling frequency of a signal that it will not distort its underlying information, should be twice the frequency of the signal's highest frequency component.

If f_S is the sampling frequency, then the Nyquist limit, f_N is defined as equal to $f_S/2$.

The next stage is improving the detection accuracy of phones by increasing the magnitude of energy in the higher frequency which improves the model's performance. First order high-pass filter is used for pre-emphasis of the signals.

Speech analysis can be based on either spectral features or temporal features [21]. Audio signals can be mathematically expressed in two domains. In the time domain, sound is expressed as a series of pressure variations that occur over a period of time. In the frequency domain, the spectrum describes audio signal in terms of the frequency components that constitute the sound. Transforms can be used to convert a specific signal between the frequency and time domains. A Fourier transform converts a time-based signal into an integral of various frequency features like fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. and the inverse Fourier transform converts the frequency-domain function back to the time function.

6. FEATURE EXTRACTION USED IN MALAYALAM ASR

The speech feature extraction is a categorization problem. It reduces the dimensionality of the input vector without losing the discriminating power of the signal. The objective of modeling technique is to generate models using feature vector. The speech recognition is divided into speaker dependent and speaker independent recognition. In the speaker independent mode of the speech recognition system ignore the speaker specific characteristics of the speech signal and extract the only intended word or message. But speaker dependent recognition system uses the speaker characteristics in the speech signal [22].

Some of the feature extraction techniques use in Malayalam ASR are discussed below.

Perceptual Linear prediction: PLP models the human voice based on the concept of psychophysics of hearing [23]. Perceptual Linear prediction improves speech recognition rate by eliminating irrelevant components of speech. In this technique the signal is windowed and the power spectrum is calculated. Bark filter bank is applied followed by equal loudness pre-emphasis in order to simulate the sensitivity of hearing. These values are transformed and processed by linear prediction. The predictor coefficients of a signal computed by LP are then used to obtain cepstral coefficients.

Wavelet Extraction: A Wavelet is a wave-like oscillation with limited duration and having two fundamental properties scale and location. Wavelet transform decomposes a signal into a set of wavelets using a mother wavelet by changing the scale and location so that the signal is represented in terms of frequency and time. In Discrete Wavelet Transform, sound signal is passed from low-pass and high-pass filter, the outputs of the filters are known as *approximation* and *detailed* coefficients respectively [24]. The approximation coefficients characterize a sound signal as compared to detailed coefficients since low-frequency components characterizes the sound signals.

Support Vector Machine

SVM is a binary classifier, which can also be used in multiclass problems. SVM is a supervised machine learning algorithm which maps the input patterns into higher dimension feature space by a nonlinear transform. SVM algorithm finds the optimal hyper plane whose distance from it to the nearest data point on each side is maximized in the higher dimension feature space which separates the data points into different classes.

Mel Frequency Cepstral Coefficients:

MFCC is a feature extraction method is the commonly used technique for speech feature extraction in Malayalam Language. MFCCs coefficients extracts the structure of the power spectrum of an audio signal. The voice signal is fragmented into different segments with each segment having 25ms width and with the signal at 10ms apart. Using a

rectangular window to segments the signal results in serious side lobe leakage will produce noise in the high-frequency domain. So a Hanning (considerably large main-lobe width) or Hamming (better in cancelling the nearest side lobe) is used to segment the signal to avoid the unwanted frequencies in the high-frequency region. The Discrete Fourier Transform (DFT) is used to derive coefficients by converting the speech waves from time domain into a frequency domain.

Our ears have can easily differentiate at a lower frequency than at a higher frequency. The Mel Scale is be used for mapping the original audio signal frequencies to the frequencies that human beings will perceive. It is followed by the application of log to the output of Mel-filter to mimic the human hearing system. The application of the mel-scale approximates the human auditory perception of sound frequency. The formula for the mapping is given below.

$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

The Mel Frequency Cepstral Coefficients are calculated from the mel-scaled spectrum. MFCCs extract those features of the speech signal that are important for human audio perception and discard less relevant information from the sample [25].

After the Discrete Fourier Transforms, the periods in the time domain and frequency domain are inverted, so that the fundamental frequency with the lowest frequency in frequency domain will have the highest frequency in the time domain. So, the inverse transform of the output from the previous step is taken. The Inverse Fourier Transform of the log of the magnitude of the signal is called cepstral coefficient or cepstrum.

The MFCC feature extraction technique takes the first 12 coefficients of the signal after applying the Inverse Discrete Fourier Transform operations. Along with the 12 coefficients, MFCC technique takes the energy of the signal sample as a feature which helps in identifying the phones. The MFCC technique also take additional 26 dynamic features such as the first order derivatives and second order derivatives of the features. MFCC technique will thus generate 39 features from each speech signal sample which are used as the input for the speech recognition model.

CONCLUSION

Only a few works are done for Malayalam ASR since last decade. In this paper, we discussed brief history of ASR systems, different feature extraction and modeling techniques used for the Automatic Speech Recognition in Malayalam language. Each feature extraction technique performs better for a particular dataset. From Table 1, we can see that MFCC is the most used method for extracting features from isolated words in Malayalam. ASR models so far developed in Malayalam are suitable for limited vocabulary speech recognition and could be improved by building ASR models for larger vocabulary.

REFERENCES

[1] J. Naga Padmaja, R. Rajeswara Rao, A Comparative Study of Silence and Non Silence Regions Of Speech Signal Using Prosody Features For Emotion Recognition, Indian Journal of Computer Science and Engineering Vol. 7, No. 4, pp. 153-161,2016.

[2] Rashmi C R ,Review of Algorithms and Applications in Speech Recognition System, International Journal of Computer Science and Information Technologies, Vol. 5 , No.4 , pp. 5258-5262, 2014.

[3] A Review on Speech Recognition Technique, International Journal of Computer Applications, Vol 10, No.3, pp. 16-24. 2010.

[4] Mohamed Hashim Changrampadi, A. Shahina, M. Badri Narayanan, A. Nayeemulla Khan, End-to-End Speech Recognition of Tamil Language, Intelligent Automation & Soft Computing,Vol. 32, No.2, pp. 1309-1323,2022.

[5] Manohar K, Jayan A R and Rajan R,Mlphon: A multifunctional grapheme-phoneme conversion tool using finite state transducers. IEEE Access, Vol.10, pp. 97555–97575,2022

[6] Ravindra Parshuram Bachate, Ashok Sharma, Automatic Speech Recognition Systems for Regional Languages In India, International Journal of Recent Technology And Engineering, Volume-8, Issue-2S3, pp. 585-592, 2019.

[7] "SHARIKA - Malayalam Speech Recognition System", Shyam.K, ICIST, 2007.

[8] Cini Kurian, Kannan Balakrishnan, Malayalam Isolated Digit Recognition using HMM and PLP Cepstral coefficient", International Journal of Advanced Information Technology, Vol. 1, No.5, pp.31-38,2011.

[9] Cini Kurian, Kannan Balakrishnan, Development& Evaluation of Different Acoustic Model for Malayalam Continuous Speech Recognition, International Conference on Communication Technology and System Design,2011.

[10] CiniKurian, Kannan Balakrishnan, Connected Digit Speech Recognition System for Language", Indian Academy of Sciences,Sadhana, Part 6, Vol.38, pp. 1339–1346, 2013.

[11] Sonia Sunny, David Peter S, K Poulose Jacob,Development of a Speech Recognition System for Speaker Independent Isolated Malayalam Words·International Journal of Computer Science & Engineering Technology

[12] Lekshmi.K.R, Elizabeth Sherly,An Acoustic Model for Isolated Malayalam Speech Recognition with different Gaussian Mixtures(7), National Conference on Indian Language, Cochin University of Science and Technology, 2018

[13] Lekshmi K R and Sherly E, An ASR system for Malayalam short stories using deep neural network in Kaldi. International Conference of Artificial. Intelligence and. Smart Systems, pp. 972–979, 2021.

[14] K R Lekshmi and Elizabeth Sherly, An Acoustic Model and Linguistic Analysis for Malayalam Disyllabic Words: A Low Resource Language, International Journal of Speech Technology, Vol24(2), No.10, pp. 483-495, 2021.

[15] JasminSashish ,Abraham Samuel, Rajeev Rajan, AStudy On Conventional And Syllable-Based

Approaches For Automatic Speech Recognition In Malayalam, Sadhana, Vol, 47, No,284,pp.1-5,2022.
[16] Arun H P, Jithin Kunjumon, Sambhunath, Ancy S Ansalem, Malayalam Speech to Text Conversion Using Deep Learning, IOSR Journal of Engineering, Vol. 11, No. 7, pp. 2278-8719, 2021
[17] V K Muneer, K P Mohamed Basheer, Rizwana Kallooravi Thandil, Convolutional Neural Network-Based Automatic Speech Emotion Recognition System for Malayalam, Indian Journal of Science and Technology, Vol. 16, No.46, pp.4410-4420, 2022.
[18] Kavya Manohar, Jayan A R and Rajeev Rajan, Improving Speech Recognition Systems for The Morphologically Complex Malayalam Language Using Subword Tokens For Language Modeling, EURASIP Journal On Audio, Speech and Music Processing, Vol.2023, No.47, pp. 2023:47, 2023.
[19] Leena G Pillai, D Muhammad Noorul Mubarak, Malayalam language vowel classification using Support Vector Machine for children, Sadhana, The Indian Academy of Sciences, Volume 48, No. 41, pp. 1-10,2023.
[20] Sania Iqbal, Speech Recognition Systems – A Review, International Journal of Advance Research in Science and Engineering, Vol. 07, Special issue No.04, pp.2046-2059, 2018.
[21] Constantin Constantinescu, Remus Brad, An Overview on Sound Features in Time and Frequency Domain, International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, Vol. XIII, No.1, pp 45-58, 2023.
[22] Manoj Kumar Sharma, Omendri kumara, Speech Recognition: A Review, National Conference on Cloud Computing & Big Data, 62-71
[23] Namrata Dave, Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition, International Journal for Advance Research In Engineering And Technology, Vol 1, Issue VI, pp.1-5,2013.
[24] Pardeep Sangwan, Dinesh Sheoran, and Saurabh Bhardwaj, Speech Recognition using Wavelet based Feature Extraction Techniques ,Global Journal of Enterprise Information System, Vol. 9, Issue 2, pp. 23-27, 2017.
[25] Maria Labied, Abdessamad Belangour, Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison, International Journal of Advanced Computer Science and Applications, Vol. 12, No. 8, pp.177-182, 2021.