Automatic Speech Recognition System: Automatic Speech to Text Using Acoustic Modelling and Deep Learning

Dr.R. Jayalakshmi[1], Sagar D[2], Somalaraju Yashwanth Varma[3], Syed Mohammad Musharraf[4], Tejaswin R M[5], N Venkata Vamshi[6], Yelugubanti Manikanta[7]

[1]Guide, Asst. Prof ECE Dept.

[3,4,5,6,7]Dept. of Electronics & Communication Engineering, Sri Chandrasekharendra Saraswati Viswa Mahavidyalaya, Enathur, Kanchipuram - 631 561

Abstract-Language is the most important means of communication and speech is its main medium. Many research activities are being conducted on Automatic Speech Recognition. But ASR systems have a major drawback in their performance i.e., efficiency. Improving the efficiency in an ASR system is quite difficult. Currently, research is being carried out to the finding of the next state of the world using Hidden Markow Model (HMM). Our Study concludes that for ASR systems, Deep Learning techniques is a more suitable application, because it increases the efficiency of the whole process. We are going to represent our work on building a speaker independent, large vocabulary continuous speech recognition system for English and Hindi. Our Study concludes that for ASR systems, Deep Learning techniques is a more suitable application, because it increases the efficiency of the whole process. Based on the conclusion, we will utilize readily available language models to build an offline live Speech recognition to Text System and Translation. Using offline speech recognition toolkits like VOSK and Kaldi an offline model is developed. Argos translate an open-source library is used to translate English speech to Hindi text. Raspberry Pi 4 is used to implement the offline speech recognition module

Keywords: Speech recognition system, speech processing, Feature extraction techniques, modelling techniques, applications of SRS, NLP and ASR system, Word Error Rate(WER), CNN and RNN .

INTRODUCTION

In recent decades, researchers have been increasingly interested in automatic speech recognition (ASR) since speech is a method of communication between people. Automatic Speech Recognition (ASR) is a process in which human speech is converted into text. *ASR is being widely used to development of smart devices, smart voice user interfaces, applications of people with disabilities, military, robotics and various fields.* ASR began with simple systems that responded to a limited number of sounds and has evolved into sophisticated systems that respond fluently to natural language. Because of the desire to automate simple tasks that require human-machine interaction, there has been increasing interest in ASR technology [2]. ASR can be defined as the process of deriving the transcription of speech, known as a word sequence, in which the focus is on the shape of the speech wave. In actuality, speech recognition is difficult because of the diversity in speech signals. Currently, ASR is widely applied in many functions, such as weather reports, automatic call handling, stock quotes, and inquiry systems. Communication can be divided into human-human communication and human-machine communication. Human-to-human communication may be limited depending on the language used, as speakers may need a third party to translate speech, such as in unified messaging systems. More recently, human-machine communication has improved greatly by using speech techniques, for example, voice search, games, and interaction systems in the context of a household living room. According to, ASR studies are affected by the following:

- Number of Speakers. To train a system, speech from a large number of users is needed.
- Nature of the Speech. The user’s voice is more easily recognized in an isolated recognition system by having the speech uttered word for word with pauses in between them.

- Vocabulary Size. Speech recognition systems vary based on the number of words that they can recognize.
- Spectral Bandwidth. If bandwidth decreases, the performance of the trained ASR system will be poor, and vice versa.

  In this research, we aim to help to increase the efficiency and to decrease the word error rate using deep learning technique. Speech recognition is a technique that can transcribe user speech in text format. And text translation converts the text in other languages. Both recognizer and translator are available separately that work in offline. Using deep learning techniques to build a speech recognition and translation system that works in offline. It takes the speech input and transcribes in its own language and also gives the translated output.

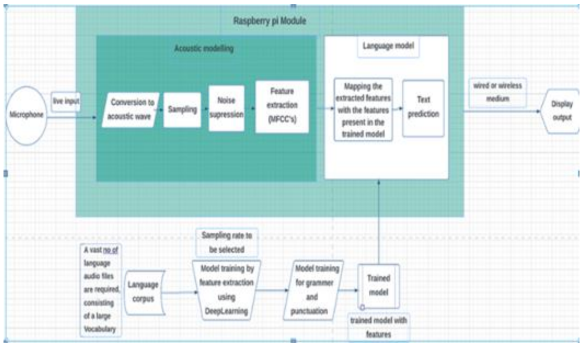
Objective:

Our Aim is to build a system that can recognize and translate in to text the Sanskrit language. Technically, we are also aiming to improve the phoneme level organizer using a variety of features, fusion of features sets, improvement in pronunciation etc., The main objective of feature extraction is to identify the discriminative and robust features in the acoustic data. To make a system that runs in raspberry pi4 that translates the speech especially for seminar halls.

METHODOLOGY

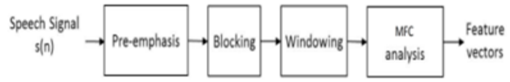
In section, the method used to conduct this research is described First, the research questions are described, followed by the search strategy. The inclusion criteria are then stated; finally, the quality assessment process is presented. The systematic review of the literature was conducted by applying the preferred reporting items. According to the flow diagram the input signal undergoes into the different modellings for the sampling and feature extraction for mapping with corpus.

ACOUSTIC MODELLLING


Raspberry pi Module
Acoustic modelling
Language model
Microphone
live input
Conversion to acoustic wave
Sampling
Noise supression
Feature extraction (MFCC's)
Mapping the extracted features with the features present in the trained model
Text prediction
wired or wireless medium
Display output
Sampling rate to be selected
A vast no of language audio files are required, consisting of a large Vocabulary
Language corpus
Model training by feature extraction using DeepLearning
Model training for grammer and punctuation
Trained model
trained model with features

Acoustic Modeling is a process of audio signal processing where the signal undergo sampling to find the letter or word in a sample. It converts acoustic analog signal to digital signal. It cuts signal into 25ms samples 10ms gap each and amplifies the power of the signal. Hanning window is applied on the signal then the DFT is applied on the samples. Taking buffer size as1024 bytes and frame rate is 16000 frames per buffer. Word level acoustic models are built by concatenating the phoneme level models. Each of the phonemes is represented by a left-right HMM. The number of states N in each HMM is set to 5. Out of these, the entry and exit states of the HMM are non-emitting. They act as the joining points for the phoneme HMMs to create word level HMMs. For each of the training sentences, the corresponding phoneme HMMs are concatenated to form a composite HMM. Now forward -backward algorithm is used to accumulate the statistics, such as state occupation counts, etc., for parameter estimation. When all the training data are processed, the accumulated statistics are used to re-estimate the HMM parameters.

FEATURE EXTRACTION



1)Pre-emphasis:

The amplitude of voiced speech falls off roughly at the rate of -6 dB/ octave at higher frequencies. So the high frequency components have lower magnitude than the low frequency ones. To compensate for this, pre-emphasis is applied to the speech signal prior to the spectral analysis. We use the first order high pass filter for pre-emphasis.

2) Blocking:

Blocking process divides the entire speech signal into overlapping segments called frames. Overlapping improves the correlation between the spectral estimates of successive frames. We use frame widths and frame shifts of 25 ms and 10 ms, respectively.

Filter bank analysis:

Mel frequency cepstral coefficients (MFCC) are widely used as features for speech recognition tasks. The computation of MFCCs emulates the processing of speech signal by human ear. Cochlea in the inner ear resolves frequencies nonlinearly across the audio spectrum. This nonlinear frequency resolution is achieved by Mel scale filter banks. They use triangular filters in the frequency domain, equally spaced in Mel scale.

Mel scale is defined as:

$$M(f) = 1125\ ln(1 + f\ /700\ )$$

MFCC computation emulates the frequency resolving mechanism of the inner ear. It calculates the magnitude spectrum for each frame, thereby identifying the frequencies present in the frame. The magnitude coefficients are weighted by the corresponding triangular filter gain and the accumulated result is a representative of the spectral magnitude in that filter channel.

Language modelling

N-gram models are widely used as language models in LVCSR systems. They can be estimated from a sufficiently large text corpus using relative frequency approach. Our work uses bigrams with back-off smoothing

WORD ERROR RATE

Automatic speech recognition (ASR) technology uses machines and software to identify and process spoken language. It can also be used to authenticate a person's identity by their voice. This technology has advanced significantly in recent years, but does not always yield perfect results.

In the process of recognizing speech and translating it into text form, some words may be left out or mistranslated. If you have used ASR in some capacity, you probably have encountered the phrase "word error rate" (WER).

The method for calculating basic WER is actually pretty simple. Basically, WER is the number of errors divided by the total words.

To get the WER, start by adding up the substitutions, insertions, and deletions that occur in a sequence of recognized words. Divide that number by the total number of words originally spoken. The result is the WER.

To put it in a simple formula,

Word Error Rate = (Substitutions + Insertions + Deletions) / Number of Words Spoken

- A substitution occurs when a word gets replaced (for example, "noose" is transcribed as "moose")
- An insertion is when a word is added that wasn't said (for example, "SAT" becomes "essay tea")
- A deletion happens when a word is left out of the transcript completely (for example, "turn it around" becomes "turn around")

Source of errors

WER does not account for the reason *why* errors may happen. Factors that can affect WER, without necessarily reflecting the capabilities of the ASR technology itself, include:

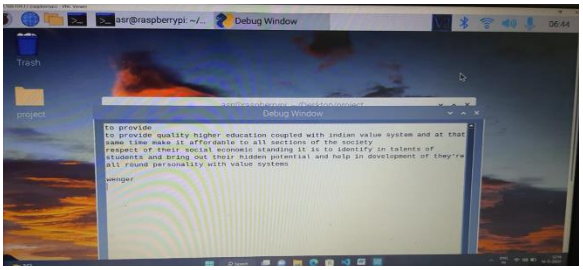
- Recording quality
- Microphone quality
- Speaker pronunciation
- Background noise
- Unusual names, locations, and other proper nouns
- Technical or industry-specific terms
- The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level.

OUTPUT FOR WORD ERROR RATE

Sample passage:

To provide "quality higher education coupled with Indian Value system and at the same time make it affordable to all sections of the society irrespective of their social or economic standing". It is to identify innate talents of students and bring out their hidden potentialities and help in development of their all-round personality with value systems.

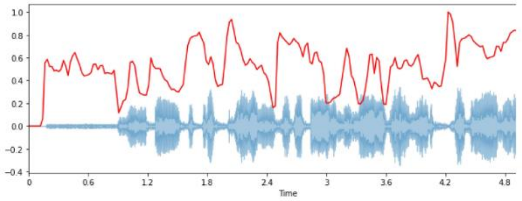
OUTPUT:



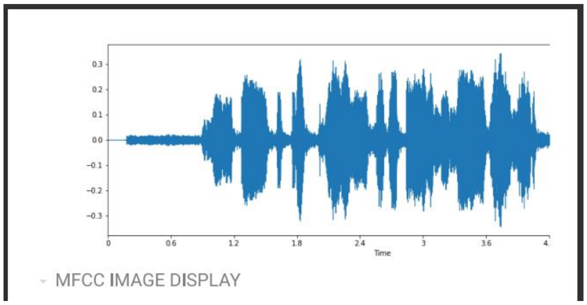
```
to provide
to provide quality higher education coupled with indian value system and at that
same time make it affordable to all sections of the society
respect of their social economic standing it is to identify in talents of
students and bring out their hidden potential and help in development of they're
all round personality with value systems

wenger
```

- We trained a sample passage with this python package and we get good efficiency and low word error rate. Due to low data set we got efficiency around 70-80%
- For display we use LCD screen display with help of raspberry pi4 and this python package files are inserted with SD card to raspberry pi4.

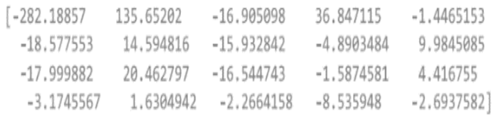
FEATURE EXTRACTION OUTPUT



Spectral Roll off



MFCC image display



```
[-282.18857   135.65202   -16.905098   36.847115   -1.4465153
  -18.577553   14.594816  -15.932842   -4.8903484   9.9845085
  -17.999882   20.462797  -16.544743   -1.5874581   4.416755
   -3.1745567   1.6304942  -2.2664158  -8.535948   -2.6937582]
```

Output values for feature extraction

Python Package:

It is a python module uploaded in python. It works as English speech to Hindi text, and it works continuously as sentence by sentence until u switch it off the command prompt. In command prompt type pip install ASRscsvmv . Then open python idle and type from ASRscsvmv import asrscsvmv. This will pop up a window so you can start speaking. Python >=3.8 is recommended.

Parameters for speech recognition:

- Number of speakers – 1 person
- Frames per buffer – 16000
- Buffer size - 1024 bytes
- Chunk – 15000bytes
- Word error rate

CONCLUSION

In this paper automatic speech recognition system for ENGLISH and HINDI is proposed. VOSK model is used to build the acoustic model and language model. Due to low dataset, we got a efficiency of nearly 75-80 percent but as it is deep learning algorithm RNN the efficiency will increase as many as the recognition takes place. The efficiency is affected by the environment also so a good environment is needed. The efficiency varies from person to person, male to female. Age also varies the efficiency as vocal card strength decreases then the efficiency decreases. Accent also varies the efficiency the limitation is this model works for only Indian English accent. Speaker's pronunciation abilities, health, communication speed, environment noise, microphone quality etc., may cause degradation of speech recognition accuracy.

REFERENCE

[1] H. Ibrahim and A. Varol, "A Study on Automatic Speech Recognition Systems," 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1-5,doi:10.1109/ISDFS49300.2020.9116286.

[2] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech recognition of Moroccan dialect using hidden Markov models," in Procedia Computer Science, Jan. 2019, vol. 151, pp. 985–991, doi:10.1016/j.procs.2019.04.138..

[3] Y. Kumar and N. Singh, "A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature Review," 2019 International Conference on Automation, Computational and Technology Management

(ICACTM), 2019, pp. 168-173, doi: 10.1109/ICACTM.2019.8776714.

[4] Madhavaraj A., Ramakrishnan A. G., “Data-pooling and Multi-Task Learning for Enhanced Performance of Speech Recognition Systems in Multiple Low Resourced Languages”, Proc. 25th National Conference on Communications (NCC 2019), 2019

[5] S. Hase and S. Nimbhore, "Speech Recognition: A Concise Significance," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), 2021.

[6] Kamini Sabu, Syomantak Chaudhuri, Preeti Rao, Mahesh Patil “An Optimized Signal Processing Pipeline for Syllable Detection and Speech Rate Estimation". Department of Electrical Engineering Indian Institute of Technology Bombay Mumbai, India, 2021.

[7] Y. Liu and K. Kirchhoff, “Graph-based semisupervised learning for acoustic modeling in automatic speech recognition,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 11, pp. 1946–1956, Nov. 2016, doi: 10.1109/TASLP.2016.2593800.

[8] J. Drexler and J. Glass, ‘‘Subword regularization and beam search decoding for end-to-end automatic speech recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019,pp. 6266–6270