International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

A Study on Automatic Speech Recognition

Nikhil Balkrishna Nair[1], Prof. Lalita Moharkar[2], Preetam Lobo[3] , Shruti Kharinta[4], Sumit Patil[5]

[1]Student, Dept. of Electronics and Telecommunication, Xavier Institute of Engineering, Mumbai, Maharashtra, India
[2]Assistant Professor, Dept. of Electronics and Telecommunication, Xavier Institute of Engineering, Mumbai, Maharashtra, India
[3,4,5]Student, Dept. of Electronics and Telecommunication, Xavier Institute of Engineering, Mumbai, Maharashtra, India

---

Abstract - *Speech recognition systems are the newest trending technologies in the world. It ranges from medical applications to automated interaction systems. It is used very significantly. In recent years various algorithms have been developed based on Recurrent Neural Networks to get better results. Due to variations in recording devices, speakers, contexts, and the environment speech recognition has become a complex task. This paper provides a study on the language models, acoustic models, and various feature extraction methods used in the automatic speech recognition system.*

Key Words: **Speech Recognition, MFCC, RNN, HMM, LSTM**

1. INTRODUCTION

Speech recognition technology enables computers to take spoken audio, then processed it into an electrical signal. Speech recognition sometimes is also referred as Automatic Speech Recognition (ASR). The signal is then converted using modern signal processing technologies, separating syllables and words. Over time, the computer can learn to understand speech from training them with data, thanks to remarkable recent advances in DeepSpeech. Presently computers render it and generate text. Still, how do computers understand human speech? Speech is just a sequence of sound waves created by our vocal cords when they cause the air to vibrate around them. These sound waves are recorded by a microphone, replacing humans in many fields. Also, artificial intelligence subparts such as machine learning, deep learning and computer vision have gone under a rapid change in the past few years. One such change is used in speech processing.

Deep neural networks have become a popular approach in Automatic Speech Recognition (ASR) systems that combines good acoustic with a language model, but they cannot handle speech data well as they don't have a way to feed information to the previous layer. Thus, RNN has been introduced to take this drawback into account. However, RNN has disadvantages like vanishing and exploding gradients. Long Short-Term Memory (LSTM) was introduced to overcome this, which is a particular type of RNN. Using LSTM in Speech Recognition has significantly achieved the best word error rate. MFCC, a pre-processing technique that helps to convert the datasets containing audio and transcripts into machine understandable language.

2. SPEECH RECOGNITION PROCESS

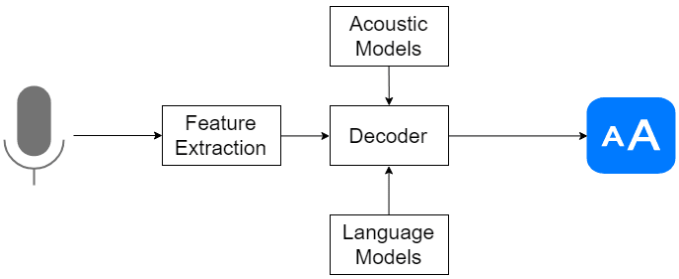

Fig-1: Block Diagram of Speech Recognition system

Speech Recognition provides how computers can be upgraded to accept speech or human voice assist input instead of giving input by keyboard. It is highly advantageous for disabled people.

The primary method of speech processing is the process of studying the speech signals and the methods of processing these signals. In the conventional speech recognition method, each word must be expressed through a feature vector and statistically available vector pattern matching using a neural network [1]. On the contrary to the antediluvian method HMM, neural networks do not require prior knowledge of speech process and do not need statistics of speech data Abbreviations and Acronyms.

2.1 Types of Speech Recognition

Depending on the type of words that the speech recognition system can recognize, the speech recognition system can be divided into the following categories:

- Isolated Word
- Continuous speech
- Connected Word
- Spontaneous speech

2.2 Types of speaker models

Based on the speaker model, speech recognition systems can be broadly divided into two main categories: speaker-dependent and speaker-independent.

1. Speaker dependent model: This system is designed for specific speakers. They are easier to develop and more accurate, but they are not very flexible.

2. Speaker independent models: These systems are designed for a variety of speakers. These systems are challenging to develop and less accurate, but they are very much flexible.

2.3 Types of vocabulary

The vocabulary size of an ASR system affects processing requirements, complexity, and precision of the system. Speech Recognition System: In speech-to-text, types of vocabulary can be classified as follows:

1. Small vocabulary: single letter.
2. Medium vocabulary: words consisting two or three letters.
3. Large vocabulary: more letter word [2]

2.4 Speech Pre-Processing

It plays an influential role in canceling out the trivial sources of variation. The speech preprocessing generally includes windowing, noise filtering, reverberation canceling, and smoothing, all of which helps to improve speech recognition accuracy

2.5 Feature Extraction

The undertaking of the acoustic front-end is to separate trademark includes out of the expressed expression. It usually takes an edge of the speech signal each 16-32 msec and refreshes each 8-16 msec, and plays out a particular examination. The regular front-end includes the following blocks: -

Fast Fourier Transformation (FFT), calculation of logarithm (LOG), the Discrete Cosine Transformation (DCT), and Linear Discriminate Analysis (LDA). It should retain helpful information of the signal, deduct redundant and unwanted information, show minor variation from one speaking environment to another, usually occur and naturally in speech.

There is a probability of identifying speech with a theoretical waveform. As a result of huge voice changes, there is an urgent need to perform some feature extractions to reduce the changes. The next section describes the extraction of some features.

Methods used are:

1. LPC (linear predictive coding)
2. MFCC (Mel frequency cepstral coefficients)
3. LPCC (linear predictive cepstral coefficients)
4. Rasta Filtering (Relative spectral)
5. PLDA (Probabilistic Linear Discriminate Analysis) [3]

Another method is based on Mel-frequency Cepstral Coefficients (MFCC). Highlight's extraction in ASR is the calculation of a grouping of highlight vectors, giving a reduced portrayal of the given speech signal. It includes 3 main stages: -

1. The principal stage is called the speech investigation or the acoustic front-end, which performs a spectra-fleeting examination of the speech signal and creates highlights portraying the envelope of the force range of short speech spans.
2. The second stage assembles an all-encompassing element vector made out of static and dynamic highlights.
3. At long last, the last stage changes these all-encompassing element vectors into more minimized and robust vectors that are at that point provided to the recognizer.

Mel Frequency Cepstrum Coefficients (MFCC): -

The Most famous element extraction methods utilized in speech acknowledgment depend on the recurrence area utilizing the Mel scale, which depends on the human ear scale and is more accurate than time-domain features. Mel-Frequency Cepstral Coefficients (MFCC) portray the genuine cepstral of a windowed short time signal got from the Fast Fourier Transform (FFT) of that signal.

MFCC is a sound element extraction method that extricates boundaries from the speech like those utilized by people for hearing speech while simultaneously deemphasizing all other data. In many frameworks covering the frames is utilized to smooth progress from casing to outline. Each period is then windowed with Hamming window to kill discontinuities at the edges. FFT is utilized to accelerate the processing. The logarithmic Mel-Scaled channel bank is applied to the Fourier changed edge. This scale is roughly direct up to 1 kHz and logarithmic at more prominent frequencies. The connection between recurrence of speech and Mel scale can be set up as:

Recurrence (Mel Scaled) = [2595log (1+f (Hz)/700]

MFCCs use Mel-scale channel bank where the higher recurrence channels have more impressive transfer speed than the lower recurrence channels; however, their worldly goals are the same.

For every speech outline, a bunch of MFCC is registered. This arrangement of coefficients is called an acoustic vector which speaks to the phonetically significant qualities of speech and is extremely valuable for additional investigation and handling in Speech Recognition. We can utilize the initial 20 to 40 edges that give an excellent assessment of the speech. All out of 42 MFCC boundaries incorporate twelve unique, twelve delta, twelve delta-delta, three log energy, and three 0th boundary.[5]

Body Linear Predictive Codes (LPC): -

It is helpful to pack signals for productive transmission and capacity. A computerized signal is compacted before transmission for productive use of channels on remote media. For medium or low piece rate coders, LPC is most

broadly utilized. LPC is one of the most remarkable speech analysis procedures, and it has picked up fame as a formant assessment strategy. Rather than moving the whole signal, we can move this leftover mistake and speech boundaries to produce the first signal. A parametric model is figured dependent on the least mean squared error theory. This procedure is known as straight expectation (LP). By this strategy, the speech signal is approximated as a straight blend of its p past examples. In this strategy, the got LPC coefficients depict the formants.
In this way, with this strategy, the areas of the formants in a speech signal are assessed by registering the straight predictive coefficients over a sliding window and finding the tops in the range of the subsequent LP channel. In a speech, during vowel sound, vocal strings vibrate pleasingly; Thus, semi-intermittent signals are delivered. While in the event of consonant, excitation source can be considered as random noise. The vocal parcel functions as a channel, which is answerable for speech reaction. The phenomenon of speech generation can be effectively changed over into an equal mechanical model. The intermittent drive train and arbitrary noise can be considered as the excitation source and advanced channel as a vocal lot.[5]

Perceptual Linear Prediction (PLP): -
PLP disposes of unessential data of the speech and accordingly improves speech acknowledgment rate. PLP is indistinguishable from LPC. Aside from that, its attributes have been changed to coordinate qualities of the human hear-able framework. PLP approximates three fundamental perceptual angles to be specific: the primary band goal bends, the equal-loudness bend, and the force clamor power-law connection, which are known as the cubic-root.

The initial step is transforming from frequency to bark, an excellent portrayal of the human hearing goal in frequency. The hearable twisted range is tangled with the power range of the reenacted basic band concealing bend to reproduce the essential band incorporation of human hearing. The three stages of frequency twisting, smoothing and inspecting coordinates into a single channel bank called Bark channel bank.[5]

2.6 Acoustic modeling

It is the fundamental part of the ASR system. In acoustic modeling, the connections between the information and phonemes of the audio signal are established. The acoustic model plays a vital role in the performance of the system and responsible for the computational load. Training establishes co-relation between the introductory speech units and the acoustic observations. Training of the system requires creating a pattern representative for class features using one or more patterns that correspond to speech sounds of the same class. Many models are available for acoustic modeling out of them, Hidden Markov Model (HMM) is widely used and accepted as it is an efficient algorithm for training.

2.7 Language model

It has structural constraints available in the language to give the probabilities of occurrence. It induces the probability of a word occurrence with respect to the word sequence. Each language has its own constraints. In speech recognition, the computer system matches these sounds with word sequence. The language model distinguishes words and phrases that have similar sounds. For example, in American English, phrases like "recognize speech" and "wreck a nice beach" have the same pronunciation but mean very different things. These ambiguities are easier to resolve when the evidence of the language model integrates with the pronunciation and acoustic models.[4]

2.8 Pattern Classification

Pattern Classification (or recognition) compares the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. In testing, after completing the system training, the patterns are categorized to recognize speech.[4]

3. Recurrent Neural Network

Recurrent Neural Network is a feedforward neural network with internal memory. The RNN performs the same function for all inputs of the data, while the current input output is essentially recurrent because it relies on the last one calculation. After generating the output, it is copied and sent to the recurrent network. The decision is made taking into account the current input and the output learned from the previous input.

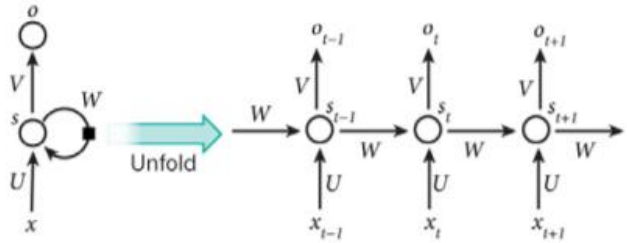

Fig-2: RNN

Unlike feedforward neural networks, it uses the RNN internal memory state to process the input sequence. Therefore, it can be applied to tasks such as undivided connected handwriting recognition or speech recognition. In other neural networks, all inputs are not dependent on each other. However, with RNN, all inputs are correlated.

RNNs use the idea of sequential information. RNN is a neural network whose memory influences future predictions. Sequential information is stored in the memory of RNNs used for predictions. The idea to use RNN instead of the traditional neural network is that in the traditional neural network, it is assumed that every input & every output does not depend on each other. Hence using a

traditional neural network is a bad idea in speech processing.[3] Prediction of any words in a sentence requires the information about the word which is utilized before, i.e., the past word which is processed. Having a memory is one of the specialties of RNN that makes it unique from other networks. There are a variety of neural networks that can be used between them. Recurrent Neural Network [RNN] is used in speech recognition to be more efficient than other networks. [6]

Algorithm

Steps involved in the RNN algorithm are: Xt is input at time t, Xt-1 is past input, and Xt+1 is the future input (sampled sound) II. St is the hidden state. It is the hidden memory. St is calculated as: St = f(U*Xt + W*Xt-1). Ot is output at step t. For example, if we want to predict the next word in a sentence, it Would be a vector of probabilities across our vocabulary, Ot=softmax(V*St).

Few things to note here are: State St is the memory of the recurrent neural network that can be hidden. St stores the data of what things took place in all the previous or past time steps. Output at step Ot is calculated exclusively based on the memory at the time "t". As mentioned above, it's a little more complicated in practice and practical implementation because St normally can't capture data from too many time steps ago.

For implementation, a traditional neural network that is deep uses various parameters at every layer while RNN shares the same parameters. It uses (U, V, W) parameters as shown by all steps above. This shows us that we are doing the same task at every single step, by passing various inputs at a different step. There is a decrease in the number of parameters in all that needs to be learned.

The diagram shown above has outputs at each time step, but it is not necessary depending on the task to perform. Consider an example where we have to predict the sentiment of a sentence. Here we only have a concern about the final output, and not the sentiment or the output which is given after each word. The Same case for inputs too, we don't need inputs at each time step. A prominent feature of an RNN is its hidden state, which stores some data about a sequence.[6]

4. LSTM

The most commonly used RNN is LSTM and is shown in Fig 3. Long and short-term memory (LSTM) is a multi-layered unit of an RNN. As you know, RNN faces a problem of long-term dependencies, which can be eliminated using a new form of RNN called LSTM. All RNN is formed consisting of a repeating structure, but in the case of LSTM, the structure varies a bit. It consists of four structure, unlike RNN which have only one. The main element in LSTM is the cell state which makes any information flow through it. Also, we can add or remove any information as per requirement using gates. There are 3 gates in LSTM which are input gate, output gate, and forget gate. The functions of these gates are to protect or control cell state.

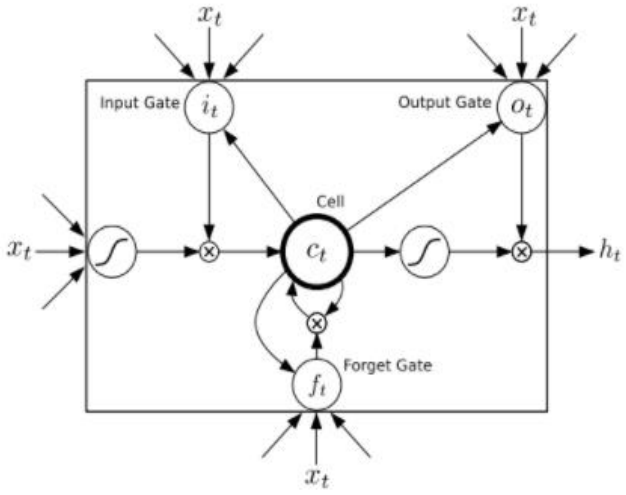

Fig-3: Long short-term memory cell

LSTM network consists of a sigmoid function with only two values at its output; either it will pass everything at the input, or it won't. Both cases are possible. So, using a cell state, we can control the long-term dependencies that were causing problems in the case of RNN. Hence LSTM finds its way to be utilized in newer versions of speech recognition software.[7]

5. Conclusion

Thus, in this paper, we have reviewed different techniques and approaches that are used to perform the task of speech recognition. Also, we have discussed the basic architecture of an ASR, it can be concluded that speech recognition is mainly dependent on the feature extraction, acoustic model, and language model.

REFERENCES

[1] Vansantha Kumari, G.Vani, Dr. R. L. K. Vankateswarlu, Dr. R. Jayasar, "Speech Recognition by using Recurrent Neural Network" IJSER volume.2, issue.6, June 2011.

[2] Prasad, V. (2015). Voice recognition system: Speech-to-text. Journal of Applied and Fundamental Sciences, 1(2), 191.

[3] Sanket Shah, Hardik Dudhrejia, 0, Speech Recognition using Neural Networks, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 07, Issue 10 (October – 2018),

[4] Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A review on different approaches for speech recognition system. International Journal of Computer Applications, 115(22).

[5] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), 1-4.

[6] Amberkar, A., Awasarmol, P., Deshmukh, G., & Dave, P. (2018, March). Speech recognition using recurrent

neural networks. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-4). IEEE.

[7] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.