International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Review: Automatic Speech Recognition

Chethan S[1], Amrutha C Howale[2], Tejas M Devang[3], Chaitra D[4]

[1-3]8th Semester Students, Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India
[4]8th Semester Student, Dept. of Electronics Communication and Engineering, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India

---

Abstract - *Language is the most important means of communication and speech is its main medium. Automatic Speech Recognition is a technology that allows human beings to speak with a computer interface in a way that its sophisticated variation, resembles normal human conversion. It is the process and the related technology for converting the speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms implemented in a device, a computer, or computer clusters. Auditory has become the primary mode of communication among human beings. However, human-machine communication is engaged more towards living with the limitation of input or output devices rather than the convenience of humans. It would be great if the computer could listen to human speech and commands. In human to machine interface, speech signal is transformed into analog and digital wave form which can be understood by machine. The paper gives the overview on production of speech sound, basic building blocks of speech processing, its application, approaches and also different approaches which are used for speech recognition system.*

Key Words: **Speech Recognition, Hidden Markov Model, Acoustic Modelling, Deep Neural Network, AISonic**

1. INTRODUCTION

Designing a machine that converse with human, particularly responding properly to spoken language has intruded engineers and scientist for centuries. Automatic Speech Recognition is the process of deriving the word sequence of a given speech waveform. Speech understanding goes to a step further in order to carry out the speakers command. The section mentions the salient application of Automatic Speech Recognition and lists the types of speech recognition system. After discussing the basic production of speech sound, variability of speech recognition that makes the speech hard, Signal processing and matching of patterns with trained models, limitations of current ASR and conclusion.

2. APPLICATIONS OF ASR

ASR Technology facilities physically handicapped persons to command and control the devices. Even to make task easier, ordinary person prefers voice commands over physical devices like keyboard or mouse. ASR technology is widely used in handheld devices like mobile and laptops. Dictation machine is the best application of ASR.

3. LITERATURE

The first machine that recognized speech was probably a toy named “Radio Rex” which came out in 1913 and was sold in the 1920’s. Basically it’s a mechanical lever action device that was a celluloid dog that moved when the spring was released by 500 Hz acoustic energy. In order to develop systems for ASR, attempts were made in 1950s where researchers studied the fundamental concepts of phonetic-acoustic. Most of the systems in 1950 were used for recognizing speech examines the vowels spectral resonances of each utterances.

Bell Laboratories designed in 1952 the "Audrey" system, which recognized digits spoken by a single voice. Ten years later, IBM in 1962 demonstrated its “Shoebox” machine, which could understand 16 words spoken in English. Fry and Denes in 1959 tried to build a recogniser with four vowels and nine consonants by using a spectrum analyser and a pattern matcher to make the recognition decision. IBM researchers studied in large vocabulary speech recognition. A large number of algorithms were used to find the number of distinct patterns.

Carnegie Mellon University’s Harphy system recognize speech with vocabulary size of 1011 words with reasonable accuracy. It was the first to make use of finite state network to reduce computation and determine the closest matching strings. In 1980, the key focus of research was on connected words speech recognition. In 1980s, Moshey J. Lasry studied speech spectrogram of letters and digits and developed a feature based speech recognition. There was a change in technology in 1980 from template based approaches to statistical modelling approach especially HMM in space research. Despite their simplicity, N-Gram language models have proved remarkably powerful. Nowadays, most practical speech recognition systems are based on statistical approach and their results with additional improvements have been made in 1990s. In 1980, Hidden Markov model (HMM) approach is one of the key technologies developed.

Neural networks to speech recognition problems is the another technology that was reintroduced in the late 1980s. A weighted HMM algorithm is proposed to address HMM based speech recognition issues of robustness and discrimination. A narrative approach for HMM speech recognition system is based on the use of a neural network as a vector quantization which is remarkable innovation in

training the neural network. For noisy environment, for robust speech recognition, a new approach to an auditory model was proposed. This approach is computationally efficient as compared with other models. In 2005, some improvements have been made on Large Vocabulary Continuous Speech recognition system for performance improvement

Deep neural network (DNN) which was introduced in 2010 which tremendously improved the performance in ASR system. Deep learning is a part of border family of machine learning methods based on artificial neural networks. DNN has many hidden layers with a large number of non-linear units and produce large number of outputs.

4. TYPES OF SPEECH SIGNALS

Speech recognition system can be categorized into different groups depending on the nature of the input.

4.1 Based on Number of Speakers:

A system is said to be speaker independent if it can recognize the speech of every speaker, such that the system has learnt the characteristics of a large number of speaker. A large number of speakers are necessary for the system to train the system. Such a system does not recognize others speech. These models are difficult to implement and are more expensive. The system is said to be dependent if it depends on a specific speaker. These types of models are accurate and less expensive

4.2 Nature of utterance:

In speech recognition system, the ability to recognize the input speech can be subdivided into different types. The user is required to provide clear pause between words in Isolated Word Recognition System. The response will be better for single word but poor result for multiple words. A Connected Word Recognition System can recognize words from small set, spoken without a pause or small duration of pause between them. Continuous Speech Recognition can recognize the speech spoken continuously. This is difficult to implement because they utilize special method of implementation. Spontaneous Speech Recognition can handle speech such as, mispronunciation, false statements, grammatical errors present in the speech

4.3 Vocabulary Size:

The accuracy, complexity and processing depends on the vocabulary of Speech Recognition System. Some applications require few words and some require large words. These type of vocabulary can be classified as, Small vocabulary, Medium vocabulary, Large and Very Large vocabulary. Small vocabulary can recognize around ten words which are used under the domain of telephone number recognition, Medium vocabulary can recognize around hundreds of words which are used under the domain of command and control sections, Large vocabulary can recognize around thousands of words which are used under the domain like dictation system respectively.

5. PRODUCTION OF SPEECH SOUND

A knowledge of generation of various speech sounds will help us to understand spectral and temporal properties of speech sounds. This, thusly, will empower us to describe sounds as far as highlights which will help in acknowledgment and classification of speech sounds.

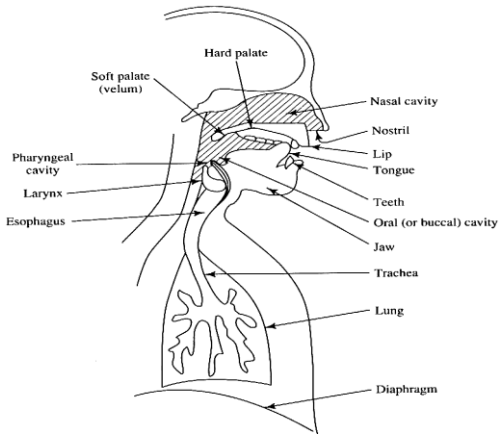

Fig -1: Human vocal system

Sounds are created when air from the lungs energize the air pit of the mouth. Figure 1 shows the human anatomy relevant to production of sound. In the event of creation of a voiced sound, say vowel [a], the glottis opens and closes periodically. Thus, puffs of air from lungs energize the oral cavity. During the conclusion times of the glottis, resonances are set up in the oral hole. The waveform leaving the lips has the mark of both the excitation and the full depression. The frequency of vibration of the glottis is prominently known as the pitch frequency.

For the creation a nasal sound, the oral section is blocked, and the velum that regularly hinders the nasal entry is lifted. During the creation of unvoiced sounds, the glottis doesn't vibrate and is open. The oral pit is energized by aperiodic source. For instance, in the creation of [s], air hurrying out of a restricted narrowing between the tongue and upper teeth energizes the cavity in front of the teeth.

In order to produce different sounds, a speaker changes the size and shape of oral cavity by movement of articulators such as tongue, jaw, lips. The full oral tract is commonly displayed as a period shifting straight filter. Such a model of discourse creation is known as a source-filter model. The excitation source can be occasional or irregular (model: [s]) or both (model: [z]).

6. HOW SPEECH IS RECOGNISED

Speech recognition is a special case of pattern recognition. Figure 2 shows the processing stages involved in speech recognition. The two stages in pattern recognition is training and testing. The process of extraction of features relevant for classification is common to both phases. During the training phase the parameters of the classification model are estimated using a large number of training data. During the testing phase, the feature of speech data is matched with the trained model of each class. The test pattern is declared to belong to that class whose model matches the test pattern the best

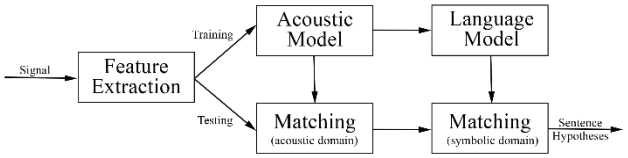

Fig -2: A block diagram of typical speech recognition

There are some important concepts in this matching process. First of all it's the concept of features. Since the quantity of parameters is enormous, we are attempting to improve it. Numbers that are calculated from speech usually by dividing the speech into frames. Then for each frame, typically of 10 milliseconds length, we extract 39 numbers that represent the speech. That's called a feature vector. The best approach to produce the quantity of parameters is a subject of dynamic examination, yet in a simple case it's a subsidiary from the spectrum.

Second, it's the concept of the model. The model of speech is called Hidden Markov Model or HMM. Hidden Markov Model (HMM) is defined as "a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobserved (hidden) states". It's a generic model that describes a black-box communication channel. In this model process is described as a sequence of states which change each other with a certain probability. This model is intended to describe any sequential process like speech. HMMs have been proven to be really practical for speech decoding.

Third, it's a matching process itself. Since it would take longer than universe existed to compare all feature vectors with all models, the search is often optimized by applying many tricks. At any points we maintain the best matching variants and extend them as time goes on, producing the best matching variants for the next frame.

7. HIDDEN MARKOV MODEL (HMM)

The task of speech recognition is to convert speech into sequence of words by a computer program. The ultimate goal of speech recognition is to enable people to communicate more naturally and effectively, while the long term objectives requires deep integration with Natural Language Processing components. Most of the modern systems are typically based on models such as Hidden Markov Models (HMMs). One reason why HMMs are popular is that their parameters can be estimated automatically from a large amount of data, and they are simple and computationally feasible. Speech system typically uses context free grammars (CFG) for the same HMMs are used for acoustic modelling.

The division of acoustic modelling and language modelling can be described as a fundamental equation of statistical speech recognition:

$$\widehat{W} = \arg\max_{W} P(\mathbf{W}|\mathbf{X}) = \arg\max_{W} P(\mathbf{X}|\mathbf{W})P(\mathbf{W})$$

where, X is the feature vector sequence, W is the corresponding word sequence that has maximum probability P(W|X). P(W) and P(W|X) constitute the probabilistic quantities computed by language by language modelling and acoustic modelling components of speech recognition systems.

For large vocabulary speech recognition, we need to decompose a word into a sub word sequence as shown in the Fig.3, since there are a large number of words.

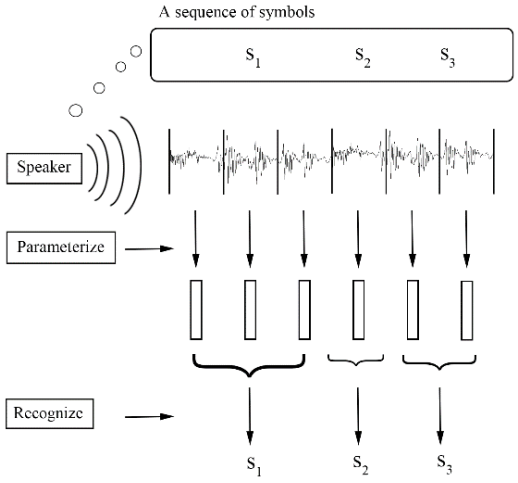

Fig -3: Message encoding and decoding.

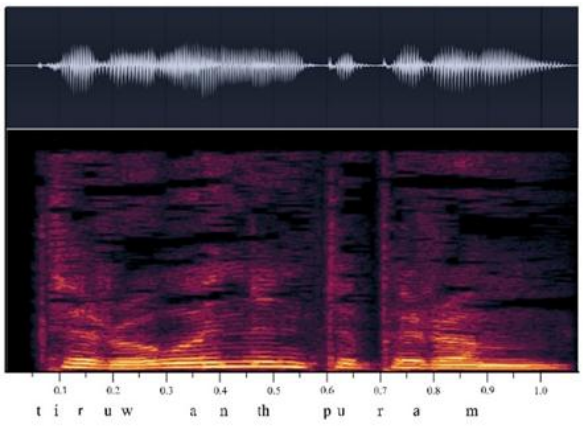

Fig -4: Message encoding and decoding.

Fig.4 shows the spectrogram of the word "Thiruvantapuram". Firstly, it is mispronounced as "Tiruvanthpuram". Also, notice the phonetic context dependent differences in formant trajectories of two occurrences of the same vowel [u]. In the first case (at 0.15-0.22sec), the second formant is decreasing whereas in the second case (at 0.62-0.67sec), it is increasing slowly.

8. KNOWLES INTELLIGENT AUDIO

The AISonic IA8201 is the industry's first mobile-centric audio edge processor specifically designed for advanced audio and machine learning applications, enabling power-efficient intelligence and privacy at the edge. It offers robust voice activation and multi-microphone audio processing with the compute power to perform advanced audio output, context awareness and gesture control for today's most advanced consumer electronics, optimized for power-sensitive applications. The IA8201 includes a high compute 128-bit core (DMX) with Knowles proprietary instruction set and a Tensilica HiFi3 core (HMD), both with Knowles audio enhancements. The DMX is a 4-way floating-point SIMD processor targeted towards efficient, high-performance computing (e.g. beam-forming, barge-in, AEC), while the HMD is targeted towards efficient, low-power, wake-on-voice applications with a two-way floating-point SIMD processor. Both cores contain dedicated accelerators for FFT, peak finding and DNNs.

A rich set of audio, and general purpose high speed interfaces enable flexible interfacing with digital microphones, other sensors, and a host for further processing. 1MB of user RAM enables storage of multiple algorithms and voice keywords

9. APPLICATIONS OF ASR

9.1 Alexa for Business:

In late 2017, Amazon reported new voice- environment, trusting that verbal orders, for example, "Alexa, print my spreadsheet," would be acquainted with supplant normal office assignments. Dubbed Alexa for Business, clients will have the option to give voice orders to start video conferences, get to schedules and print reports, and oversee individual and shared Echo gadgets, just as a large number of other regular working environment capacities. "You never again ever need to dial in a gathering ID," Amazon's Chief Technology Officer Werner Vogels stated, "Simply state 'Alexa, start the meeting." actuated instruments for the working

9.2 Microsoft Crotona:

Microsoft's Cortana has likewise started to deal with a portion of the more difficult office undertakings, for example, booking gatherings, recording meeting minutes, and making travel courses of action. Down the line, Microsoft authorities have demonstrated a gathering room where Cortana welcomes meeting members, helps them in joining a booked gathering, transcribes meeting notes, prescribes records, and helps individuals to remember the names and titles of meeting members.

Be that as it may, with Alexa and Cortana's organization (permitting Alexa clients to get to Cortana and its capacities, and the other way around) declared in the mid-year of 2017, Alexa is by all accounts moving quicker than Microsoft envisioned. The organization permits Alexa (and thus, Alexa for Business) to incorporate with Office 365 and other Microsoft programs – simply like Cortana. The usefulness, likewise with inside your home, can be gotten to through Amazon Echo speakers.

9.3 Voice Technology in Finance:

"In spite of the fact that it's still early days as far as appropriation, banks see incredible guarantee in voice-activated arrangements," reports Phil Goldstein of Biz-Tech. "Why tap on your cell phone to get your financial records balance when you can simply ask Alexa?" To the extent use cases go, banks see the incentive in voice-based banking in lessening the requirement for human client assistance agents and thus, diminishing staffing costs. Approaching your own special advanced financial associate on your cell phone could likewise support consumer loyalty and maintenance.

9.4 Marketing:

Promoting experts would profit by voice-collaborators helping them with booking internet based life posts, making reports, etc.

Voice is likewise another medium yet to be researched by publicists (outside of Podcasts and radio that is). Organizations have begun investigating manners by which this sparkling new industry can be used to help improve their clients' understanding – and obviously, more ways by which the brand can get before them all in all. Starbucks' Reorder ability for Alexa permits clients to, well, request steaming hot cups of java; Hyundai assembled a Google Home mix to begin your vehicle

10. FUTURE OF NATURAL LANGUAGE PROCESSING

According to the many market statistics, data volume is doubling every year. The vast portion of this is text data. Natural Language Processing (NLP) is a sub-branch of Data Science that attempts to extract insights from text. The major reasons behind the growth of the natural language processing market are the rising focus on an enhanced consumer experience, rapid business process automation, creation of high volumes of data, and rise in the number of contact centres. Thus, NLP plays an important role in Data Science. Industry have assumed that NLP will grow exponentially in the near future. In NLP, machines are taught to read and interpret text as humans do. NLP is set to capture the voice of the customer. With the exponential growth of multi-channel data like social or mobile data, businesses need solid technologies in place to assess and evaluate customer sentiments. A significant reward of NLP is to businesses is the concept of smart assistant, which has the potential to transform customer experience and leading customer loyalty. The smart assistance have hopefully will emerge a game changer in the future

11. CONCLUSION:

Speech recognition is growing day by day and has unlimited applications. The paper reviews the overview of speech recognition process, and its applications. It has been found that HMM is the best technique in developing language model. Speech Recognition is very fascinating problem and has attracted scientist and researchers and created a technological bang on society.

REFERENCES

[1] "Automatic Speech Recognition", Samudravijaya K, Tata Institute of Fundamental Research.

[2] "Speech and Language Processing", Daniel Jurafsky, James H. Martin. (2018).

[3] "Handbook of Natural language processing", Xuedong Huang and Li Deng, (2009).

[4] "Speech Feature Extraction Techniques", Shreya Narang, Divya Gupta, IJCSMC, Vol. 4 (2015).

[5] https://www.researchgate.net/publication/306010331_Speech_Recognition_System_-_A_Review

[6] https://www.sciencedirect.com/topics/engineering/automatic-speech-recognition

[7] "Convolutional Neural Networks for Raw Speech Recognition", Vishal Passricha and Rajesh Kumar Aggarwal.

[8] https://www.aisonic.com/

[9] https://www.dataversity.net/future-nlp-data-science/