Automatic Speech Recognition And Transliteration

Anjali Kumari
Department of Computer
Dr. D.Y. Patil Institute of Technology
Pimpri, SPPU
anjalikri0711@gmail.com

Anil Kumar Gupta
Senior Member IEEE
anilkgupta@ieee.org

Ankita Kumari
Department of Computer
Dr. D.Y. Patil Institute of Technology
Pimpri, SPPU
ankitakri1504@gmail.com

Jayshree Repale
Department of Computer
Dr. D.Y. Patil Institute of Technology
Pimpri, SPPU
jayshreerepale2000@gmail.com

Dr. Rachna Somkunwar
Associate Professor
Dr. D.Y. Patil Institute of Technology, Pimpri, Pune
rachnasomkunwar12@gmail.com

Komal Ghodke
Department of Computer
Dr. D.Y. Patil Institute of Technology
Pimpri, SPPU
komalghodke99@gmail.com

***Abstract*---- Recognition of speech is a long-standing challenge for automatic speech recognition (ASR) systems and translate speech into any target language is the second most challenging task. As we all know, communication is a weapon to conveying our thought to another person or group of person. We have only one way to express or explain our thoughts to the world is communication. Here we present a technique to train acoustic models for a target language using speech data from distinct source languages. In this approach, we are focus on how to recognize speech and translate into end-user understanding language. We centre on difficulties with ASR, basic building blocks of speech processing, feature extraction, speech recognition and performance evaluation, and translation of speech into target language like Text-To-Speech (TTS), Text-To-Text (TTT), Speech-To-Text (STT), Speech-To-Speech (STS).**

1. INTRODUCTION

Speech is one of the essential tools for communication between human to the world as we know that "communication" means exchanging ideas, message or information from one individual to another through words or language. But for successful communication, we have to know the end-user language. The main challenge is to learn all language, and this is no easy task.

To overcome this problem, Automatic Speech Recognition (ASR) comes in the picture for human needs. The issue of communication between human beings to all over the world becomes increasingly significant. Automatic Speech Recognition or ASR, as it's called in short, is the technology that allows social to use their voices to interact with a computer interface in such a way that, in its most smooth variations, resembles typical human communication. One question will come in the mind that is How Does It process? Automatic Speech recognition software works by dividing the audio of a speech into different sounds, analysing each sound, applying algorithms to find the most probable word fit in that language, and transcribing those sounds into text. After this, we have to translate the speech into user understanding language for a better flow of communication. In this application we also know about the translators Like Text-To-Speech (TTS) it takes words or sentence from the speaker or other digital device and converts them into audio, Text-To-Text (TTT) it takes words or sentence from a speaker or other digital device and converts them into text, Speech-To-Text (STT) Speech to text transformation is the process of converting spoken words into texts, Speech-To-Speech (STS)

conversion is the process of converting spoken conversation into audio.

Through automated speech recognition technology and transcription software falls short of complete human intelligence, there are many benefits of using the technology--especially in business applications. In short speech recognition and translation, applications help companies save time and financially by automated business processes and providing immediate understanding on what's happening in their phone calls because a software performs the tasks of speech acknowledgement and transcription faster and more accurately than a human can, it means it's the more cost-effective job for a person to do at the same position. It can also be a long job for a person to do the rate at which many businesses need the service performed.

2. METHODOLOGY

A. Automatic Speech Recognition:-

Automatic Speech Recognition (ASR) System: Accurately translate spoken Utterances into text (words, syllables, characters, etc.).From using ASR techniques, we can convert our raw data (Speech/voice) into any required form (Text/Speech). Nowadays, ASR has become very popular in the customer service departments of big/large corporations

For example, YouTube closed captioning, voicemail transcription, Dictation system, Siri front end, etc.

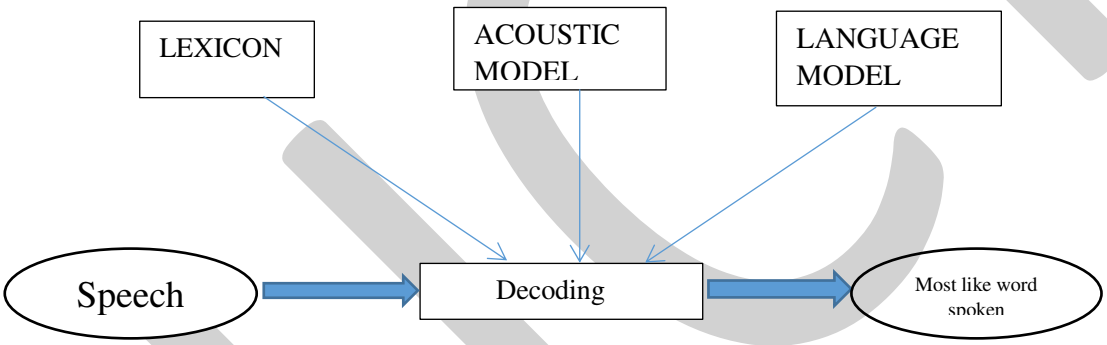

Fig.1.ASR Basic Model

i. *Lexicon*: The Lexicon plays a crucial role in automatic speech recognition as it builds the connection between the acoustic level representation and word sequence output by the speech recognizer. The lexicon specifies the sequences of phonemes which form valid words of the language.

ii. *Acoustic Model*: Acoustic model is used in ASR to represent the relationship between an audio signal and the phonemes or other linguistic (scientific study of language) units that make up speech. The acoustic model is the main component for an ASR which accounts for most of the computational load and performance of the system. The Acoustic model is being developed for detecting the spoken phoneme.

iii. *Language Model:* The role of the language model is to derive the best sentence hypothesis subject to constraints of the language. This model includes various types of linguistic information. The lexicon specifies the sequences of phonemes which form valid words of the language. The syntax

describes the rules of combining words to create proper sentences.

Let's talk about Speech, taken as an input signal and produces a modified signal by removing unwanted calls.

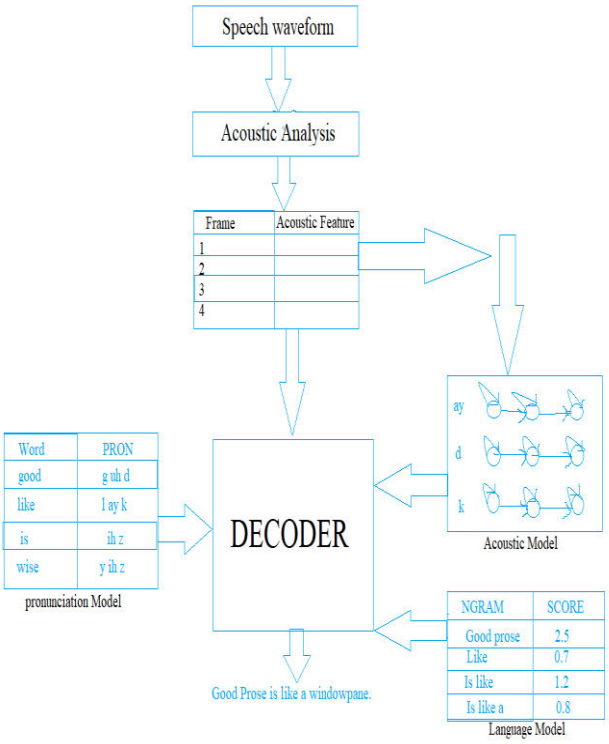

Figure.2: Structure of ASR(reference 1)

Let's move into what's the structure of a typical ASR system, so we have an acoustic analysis component which detects the speech waveform and converts it into some individual representation. So referring to the above figure2 first component is just looking at the raw speech waveform and converting it into a model which algorithm can use. We have the natural speech signal which then was discretized because we can't work with a row speech signal, sample it and generate these discrete samples, and each of these samples are typical of an order of 10 to 25milliseconds of speech and idea is that once you have each of these what we know as speech frames which are free round typically 25 milliseconds. Then we can extract acoustic features which are representative of all the information in our signal, and another reason why we discretize at particular sampling rates is that the assumption is that within each of these frames speech signal is stationary. This feature removal is a very mixed up process, and it involved a lot of signal processing, but this is also kind of motivated by how regular human ears works, actually one of the most common acoustic feature of representation which is known as Mel-Frequency cepstral coefficient or MFCCs.

Now we have these features for each frame, and these are our inputs to the next component in the ASR system, which is known is the acoustic model. There are some basic units of acoustic model: Phoneme, which is a discrete & distinctive unit of a language that can use to differentiate between words. In ASR, there is an approach which is known as the beads-on-a-string method through each word can be presented as a sequence of phonemes. Phonemes suggest the letter equal to the analogy of phonemes as letters in our written texts.

Now we move on to next model which is pronunciation model, and this is the model which provide a link between these phone sequences and words, here a typically just a simple dictionary of pronunciations is maintained. This is the only module in the ASR system that is not learned from data, the acoustic model was learned from training data, but the pronunciation model is expert derived so expertly gives us theses mappings. In this model, we have not only board sequences corresponding to the speech we also have phonetic sequences, so it's also phonetically transcribed means somebody listened to all the utterances and noted down the phone sequence corresponding to what they heard this was done by linguistic.

And the last model is what's known as the language model, in the pronunciation model the output was worded so now we have mapped a phone sequence to a particular word. Now the language model comes and explains how should these words be ordered according to a specific language, this particular model looks at lots and lots of texts in that required language, and it finds occurrences of words together.

The SRILM Toolkit is a popular Language model toolkit. It has a lot of in-built functionalities implemented; that's why it is better to use. Another toolkit which is getting quite popular nowadays is KenLM Toolkit, which handles a large volume of text very efficiently.

The ASR decoder is one of the essential components of the ASR system and has been

growing over the year to address the increasing demand for larger domains. Executing a decoder which can handle hundreds of thousands of words in the active vocabulary and hundreds of millions of n-grams in the language model in real-time is no simple task. So this is the entire scenario pipeline of how an ASR system works.

ASR desirable for:-

- Speech is the way of human communication.
- Develop natural interfaces for both literate & illiterate users.
- Contribute to the preservation of endangered language.

Difficulties with ASR:-

Several sources of variability

- Style: Conversational (or casual) speech or read the speech? Continuous speech or isolated words?
- Environment: Background noise, channel conditions, room acoustic, etc.
- Speaker characteristics: Rate of speech, accent, age (modulation), etc.
- Task specifics: Number of words in the vocabulary, language constraints, etc.

B. Translator:-

Google Translator

Introduction:-Live Speech Translation or Real-Time Speech Translation is one amongst the issue areas in speech recognition technology. Google Translate may be a complementary translation service that is developed by Google; it is a bilingual service supported neural machine artificial intelligence. The goal is to translation of text and websites from one language to different language, and conjointly for translation, it interprets multiple kinds of books and media like words, phrases and web pages. It supports Word Lens to boost the quality standard of visual and voice translation. Because it has the ability of scanning text or footage with one's device &, have it translated instantly. Moreover, the translation system works as by interpretation and automatically identifies the different languages & interprets for speech translation while not requiring somebody to faucet the button of the microphone whenever the speech translation conception is needed.

Translation methodology:

There are four main significant types of Machine translation noted as- first one is The Statistical Machine Translation [SMT], another is The Rule-Based Machine Translation [RBMT], other are Hybrid Machine Translation, & The Neural Machine Translation. Google Translator doesn't apply grammatical rules, since its algorithms depend on pattern analysis as statistics instead of traditional rule-based analysis.

i. Method of The Statistical machine translation

The Statistical Machine Translation may be a paradigm whenever is based on statistical models where one SMT Suitable for more language pairs.

ii. Google Neural Machine Translation(GNMT):

The Neural artificial intelligence is deployed by Google for a higher quality of translation. The goal is to the translation of the full phrases at a time instead of the only words, then gathering of overlapping of names for translation. Google Translator's neural artificial intelligence system improves the standard of translation over SMT in some instances. The GNMT network tries for the Interlingua artificial intelligence that encodes the "Semantics of the sentence instead of by straightforward committal to the memory of Phrase-to-Phrase Translations."

Functions:-

The Google Translator translates multiple forms of text & media, including text, speech, & text within moving or still images. Specifically, its functions include:

The Google Translator interprets multiple kinds of text & media, as well as text, speech, & text inside moving or still pictures. Specifically, its functions include:

i. *Translation of Written Words:* operate that merely interprets text or words to a requested language.

ii. *Translation of Website:* operate for translating an entire page to a specific language.

iii. *Translation of Document:* operate that interprets the document uploaded by the *user* to choose language. The documents type at ought to be within the form of: .doc, .docs, .text, .pdf, .odf, .ps, .rtf, .xl, .xls .ppt, .ppts.

iv. *Translation of Speech:* operate that interprets speech communication to the chosen language.

v. *Translation for the Mobile App:* it's known as "Tap to Translate". It makes translation instantly accessible within any app while not existing or change it.

vi. *Translation of Image:* operate that interprets text on the screen instantly by pictures & identifies text during an image taken by the user.

vii. *Written Translation*: operate that interprets language that area unit drawn on a virtual keyboard while not the support of a keyboard or written on the mobile phone's screen & because it provides the lexicon, Pronunciation and being attentive to Translation.

Comparing Machine Translation Basis methodologies:-

Period	Methodology	Description	Pros	Cons
The 1970s-1990s	Rule-Based Machine Translation (RBMT)	RBMT needs human linguistic knowledge and data effort. In contrast, particular little written resources are generally required.	i. No Bilingual Text Needed ii. Domain-Independent iii. Total Management Control iv. Reusability	i. Requires good and smart dictionaries ii. Manually set rules (Requires expertise) iii. Native translations, Language dependent Expensive to maintain, keep up and extend
The 1990s-2010s	Statistical Machine Translation (SMT)	*Core Element:* Words *Knowledge:* Phrase table *Training:* Slow Complex pipeline *Model Size:* Greater *Interpretability:* Medium *Introducing ling. Knowledge:* Possible *Open Supply Toolkit:* affirmative (Moses) *Industrial Deployment:* affirmative	i. Less manual work from linguistic consultants ii. One SMT appropriate for additional language pairs iii. Less out-of-dictionary translation: with the proper language model, the interpretation as translation is additional fluent	i. Needs bilingual corpus ii. Specific errors area units laborious to mend iii. Less appropriate for language pairs with Huge variations in word ordination iv. High Computational procedure Resources needed

2014	Neural Machine Translation (NMT)	*Core Element:* Vectors *Knowledge:* Learned weights *Training:* Slower More elegant pipeline *Model Size:* Smaller *Interpretability:* Low Opaque translation method *Introducing ling. Knowledge:* Possible (yet to be done) *Open Source Toolkit:* affirmative (Many) *Industrial Deployment:* affirmative (now at Google, Systran, WIPO)	i. End-to-End models (no pipeline of specific tasks) ii. Provides one system which will be trained to decipher the supply and target text iii. additional advanced than different strategies	i. needs Bilingual Corpus ii. Rare word drawback

Table.1: Comparing Machine Translation Basis methodologies

- Speech-to-Text:-
  - Speech-to-Text Conversion is the approach of Conversion of spoken words to written texts. The Term Recognition of Voice ought to be avoided because it is usually involved with the method of identification of an individual from their Voice. i.e. Speaker Recognition
  - *Working*:

- All Speech-to-Text systems work on a minimum of two kinds of models: An S2T acoustic Model & a Language Model, and sums up it with an extensive vocabulary system that uses pronunciation model because it is vital to know that there's no such issue as a universal speech recognizer.
- For obtaining the most superficial transcription quality, all of those models are often specialized for a given language, application domain, sort of speech and communication channels. The Accuracy of Speech Transcript is very keen about the speaker, the environmental conditions and elegance of speech. Like all different pattern recognition technology, Speech Recognition can't be error-free. It is a stricter method than what individuals ordinarily suppose, even for a personality's being. Humans area unit accustomed understanding speech, to not transcribing it and solely speech that's well developed are often transcribed with none any ambiguity. From the user's purpose of reading, a Speech-To-Text System is usually classified based mostly in its use:

Command and management, Dialog System, Text Dictation, Audio document Transcription, etc. Every use has specific necessities in terms of latency, vocabulary size, memory

constraints and editable adaptive options.

- Text-to-Speech:-
  - Text-to-Speech (TTS) will take words or sentence from the speaker or different digital device and convert them into audio. Text-to-Speech [TTS] is quite helpful technology that reads digital text aloud that's typically known as "Read Aloud" Technology.
  - *Working:*
  - TTS works with almost every personal digital device, as well as tablets and smartphones. All types of text files may be read and browse aloud, including page and word documents. Even Online website concerned pages can be read with voice.
  - The voice in TTS is a virtual voice, and reading speed will typically be sped up or delayed. Voice quality standards vary differently; however, some voices sound human. Their square measures even pc generated voices that sound like children's voice.
  - Many Tools of TTS highlight words as they're browsing aloud. This enables children to ascertain and listen to it in parallel. Some TTS tools even have a technology known as Optical Character Recognition [OCR]. That allows it to browse text aloud from pictures. E.g. If your kid could take an image of a sign with words that will turn into audio.

  - *Types of Text-to-Speech Tools:*
  - *Built-In Text-to-Speech:* Many devices have built-in TTS tools. This includes smartphones and digital tablets and chrome, Laptop computer and Desktop computers.
  - *Web-Based Tools:* Some Web site has on-the-scene TTS tools. As an example, you'll activate Website's "Reading Assist" tool, placed in this lower-left corner of your screen, to have this webpage browse aloud.
  - *Text-to-Speech apps:* You will conjointly transfer TTS Apps on smartphones and digital tablets. These Apps typically have unique options like Text lightness in several OCR and colours. Some examples embrace voice dream reader, cigar Scan Pen and workplace Lens.
  - *Chrome Tools:* Chrome could be a comparatively new atmosphere with many TTs Tools. These embraces Read browser & Write for Google Chrome and Snap that will read browser.
  - *Text-to-Speech code Programs:* There are many acquirement code programs accessible for laptop computers and Desktop Computers.

- Text to Text:-
  - *Text & Handwriting:* The Handwriting & the Text feature permits you to interprets text from one to a different language. The applying can do its best to translate to your elite language.
  - *Translate Text:* within the Google translator Text-to-Text translator works as when selecting a language to translate to be done as explicit below:
    From: At the highest left, faucet the Down arrow.
    To: At the highest Right, tap on the proper, and faucet on the Down Arrow.
  - By writing the words or phrases, you want to translate may be done. To provide the right context to your translation, write your name or phrase in a complete sentence. You'll see the Results. If not, faucet Translate sees details.

- Speech to Speech:-
  Speech-to-Speech translation systems provide a convertible atmosphere to alter verbal communication between speakers of various languages at intervals the context of specific domains.
  - *Characteristics of the Speech Recognition Module:* It's structured to produce N-Best picks and Multi-Stream Process the N-Best lists from any Speech Recognition Engines.
  - *Conversion and Voice:* The Conversations and therefore, the Voice options allow you to talk in one language, so the translation of it

another by repetition in another language.

- *Change Your Speech Settings:* Settings, as you wish in Google translator, may be applied.
- The elements of the preferred Embodiments of this inventions is consists of the following terms:
- (1)Speech recognition; (2) Machine Translation; (3) N-Best Merging Module;
(4) Verification; (5) text-to-Speech.

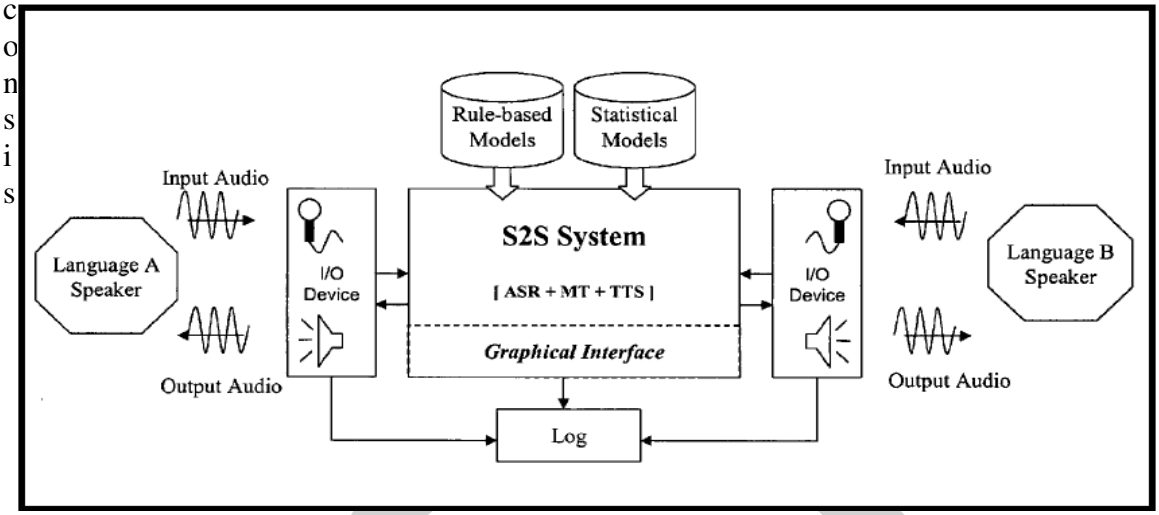

Figure.2: Speech to Speech Translation

3. LITERATURE SURVEY

Comparing different technique of ASR in Table.2 and Table.3

Period	Technique	Advantages	Disadvantages
Pre-1970	i. Audrey System	a. Audrey System built by Stephen Balashek, R.Biddulph and K. H. Davis in 1952. b. The system located the formats in the power spectrum of each utterance	single-speaker digit recognition
	ii. Shoebox Machine Speech Recognition	It was demonstrated by IBM in 1962 with 16-words machine speech recognition capability.	Word limitation
	iii. Dynamic Time Warping (DTW) algorithm	a. It is capable of operating on a 200-words vocabulary. b. It Processed speech by dividing it into short frames.	Achieving Speaker independence remained unsolved at this period.
1970-1990	i. Speech - Understanding Research (1971)	Large Vocabulary size approx 1600-words.	There is no progress in Speech recognition

	ii. Hidden Markov Model	a. Markov Chain mathematics is developed by a Leonard Baum in 1960. b. It allows the researcher to combine different sources of knowledge, such as acoustics, language, and syntax, in a unified model.	In addition to having a relationship with individual words.sequence length,word context
	iii. Tangora typewriter	a. It was created by IBM'S Fred Jelinek's in the mid-1980s. b. It can handle a 20,000-word vocabulary.	Less emphasis on match the way the human brain processes.
	iv. Dragon System	This system founded by James and Janet M. Baker. It was one of IBM'S few competitors.	Low performance. Unusual hardware configuration, Very slow and buggy,..etc

Table.2: Comparing different technique of ASR

In Practical life, actual use of Speech Recognition comes in:

Period	Technique	Advantage	Disadvantage
1987	i. Back-off-model	This model allowed language to use multiple length n-grams and Centro study Laboratory Telecommunication (CSELT) used HMM to recognize the language.	Time-consuming, it takes much time to Conversion, I.e. 100min to decode just 30sec of speech.
	ii. Recognize from Kurzweil Applied intelligent	Telecommunication (CSELT) used HMM to recognize the language.	Time-consuming,
1990	Dragon Dictate	a. A consumer product developed by Lawrence Rabiner in 1990 AT & T deployed the voice recognition call processing services in 1992 to route telephone calls without the use of the human operator. b. It has a more extensive vocabulary than the average human language.	The performance was not good; the program is prolonged and buggy.
1992	Sphinx-ii System	c. Speaker - independent d. Large vocabulary e. Continue speech recognition f. And, the best performance in DARPA, 1992 evaluation.	Word error rate is still in there.
2000's	i. GOOG-411	a. Google the first effort at speech recognition in 2007. b. GOOG-411 is 1st product on speech recognition. It is a telephone-based directory service.	Limitation in language. At that time GOOG-411support only 30 languages.
	ii. Keyword-Spotting	This technology permits analysts to search through a large volume of recorded conversation application and isolated mention of the keyword.	It is based on large vocabulary due to this spotting the right word is difficult.
	iii. Long Short-term memory(LSTM)	a. Speech recognition has been taking over by a deep learning method called LSTM.	It is quite popular but it's required deep learning .

		b. LSTM RNN's avoided the Vanishing gradient problem and learns "Very Deep Learning" tasks. c. LSTM learned by Connectionist Temporal Classification (CTC) started to outperform traditional speech recognition in a particular application.	Low performance.
	iv. Google - Speech - Recognition	a. It uses the CTC - trained LSTM and the performance jump 49%. b. It is now available through Google Voice to all smartphone users.	It required lots of memory and hardware . And its cost is high.
	v. Acoustic Modeling	a. In acoustic modelling, we use Deep feed-forward (non-recurrent) networks in 2009 by Geoffrey Hinton and other students of different university. b. The most Considerable change in accuracy since 1979. c. The applications of deep learning decreased word error rate by 30%.	Low accuracy
	vi. Gaussian Mixture Model/Hidden Markov Model (GMM/HMM)	GMM/HMM technology based on generates models of speech trained discriminatively.	Accuracy rate is low.

Table.3: Comparing different technique of ASR

4. CONCLUSION

The main aim of this paper is to make the interaction better. Automatic Speech Recognition system provides the ability to convert speech into well-understandable word. In terms of the better flow of communication ASR play an essential role in this, it takes speech as an input, and we use a translator to translate one language to another understandable language. This paper attempts to give a comprehensive survey of research in speech recognition and many different ways to translate the native language to target language. i.e
ASR supports translating audio/speech files.

5. FUTURE WORK

After this study, We need to do more to make ASR systems robust to kind variation in age , accent ,and speech ability so there are people say with speech impairment ,the distractor or they have other issues and they are not able to speak as clearly as maybe all of fast in the room. So how can we adapt ASR systems to work well with those people? And this is really, very challenging task, so how do you handle kind of noisy, real life setting with many speakers?

For e.g., Allen is sitting in a meeting, and kind of transcribing and figuring out what is going on. So ASR system have to identify the interfering speakers, this is the main speakers, this is the speaker i need to kind of transcribe .I need to haze out the other interfering speakers and so on.
If you have lots and lots of level speech. This pronunciation variability actually captured into acoustic model itself. In short we can work in future on these issues:

- Robust to variations in age, accent, and ability
- Handling noisy real-life setting with many speakers(e.g.,meeting,parties)
- Handling pronunciation variability
- Handling new language

6. REFERENCES

[1]. https://youtu.be/q67z7PTGRi8
[2]. Sadoski Furui, November 2005, 50 years of Progress in speech and Speaker Recognition Research, ECTI Transactions on Computer and Information Technology,Vol.1. No.2

[3]. International Journal of Computer Applications (0975 – 8887) Volume 41– No.8, March 2012
[4]. International Journal of Computer Applications (0975 – 8887) Volume 60– No.9, December 2012
[5]. K.H.Davis, R.Biddulph, and S.Balashek, 1952, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am.,24(6):637-642.
[6]. D.B.Fry, 1959, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299
[7]. Gerhard Rigoll, Jan.1994, "Maximum Mutual Information Neural Networks for Hybrid connectionist HMM speech Recognition Systems ", IEEE Transactions on Audio, Speech and Language processing Vol.2,No.1, PartII.
[8]. L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, and J.G.Wilpon, August 1979, Speaker Independent Recognition of Isolated Words Using Clustering Techniques , IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27:336-349
[9]. Mohamed Afify and Olivier Siohan, January 2004, Sequential Estimation With Optimal Forgetting for Robust Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol. 12, No. 1.
[10]. Mohamed Afify, Feng Liu, Hui Jiang, July 2005, A New Verification-Based Fast-Match for Large Vocabulary Continuous Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4.