ASSOSA UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

TECHNICAL REPORT

ON

AUTOMATIC SPEECH RECOGNITION

Agerie wudu

Besufekade Assefa

Bulcha Abdi

Dawit Assefa

Haile Tiruneh

Submitted to: Hawi

Submission date: January, 2017

Assosa, Ethiopia

Abstract

Speech recognition, involves capturing and digitizing the sound waves, converting them to basic language units or phonemes, constructing words from phonemes, and contextually analyzing the words to ensure correct spelling for words that sound alike. Speech Recognition is the ability of a computer to recognize general, naturally flowing utterances from a wide variety of users. It recognizes the caller's answers to move along the flow of the call. Emphasis is given on the modeling of speech units and grammar on the basis of Hidden Markov Model& Neural Networks. Speech Recognition allows you to provide input to an application with your voice. The applications and limitations on this subject enlighten the impact of speech processing in our modern technical field. While there is still much room for improvement, current speech recognition systems have remarkable performance. We are only humans, but as we develop this technology and build remarkable changes we attain certain achievements. Rather than asking what is still deficient, we ask instead what should be done to make it efficient.

Table of Contents

List of figure

Acronyms

ASR:-automatic speech recognition

SR: - speech recognition

HMM:-hidden markov models

OOV:-out of vocabulary

SRS:- speech recognition system

E.g:-example

1. Introduction

Speech recognition allows you to provide input to a system with your voice and Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words.

Just like clicking with your mouse, typing on your keyboard, or pressing a key on the phone keypad provides input to an application, speech recognition allows you to provide input by talking. In the desktop world, you need a microphone to be able to do this.

2. What is automatic speech recognition

Automatic Speech Recognition (or sometimes referred to as Speech Recognition) is the process by which a computer (or other type of machine) identifies spoken words and is the ability to translate a dictation or spoken word to text. Basically, it mean stalking to a computer & having it correctly understand what you are saying. By "understand" we mean, the application to react appropriately or to convert the input speech to another medium of conversation which is further perceivable by another application that can process it properly & provide the user the required result. The days when you had to keep staring at the computer screen and frantically hit the key or click the mouse for the computer to respond to your commands may soon be a things of past. Today we can stretch out and relax and tell your computer to do your bidding. This has been made possible by the ASR (Automatic Speech Recognition) technology.

The ASR technology would be particularly welcome by automated telephone exchange operators, doctors and lawyers, besides others whose seek freedom from tiresome conventional computer operations using keyboard and the mouse. It is suitable for applications in which computers are used to provide routine information and services. The ASR's direct speech to text dictation offers a significant advantage over traditional transcriptions. With further refinement of the technology in text will become a thing of past. ASR offers a solution to this fatigue-causing procedure by converting speech in to text.

Speech recognition is an alternative to traditional methods of interacting with a computer, such as textual input through a keyboard. An effective system can replace, or reduce the reliability on, standard keyboard and mouse input. This can especially assist the following.

- People who have little keyboard skills or experience, who are slow typists, or do not have the time or resources to develop keyboard skills.
- Dyslexic people or others who have problems with character or word use and manipulation in a textual form.
- People with physical disabilities that affect either data entry, or ability to read (and therefore check) what they have entered.

3. Automatic Speech recognition Advantages and disadvantages

3.1 Advantages Automatic Speech recognition

- **Increases productivity**

By speaking normally in to the SRS program, you create documents at the speed you can compose them in your head. People without strong typing skills or those who don't wish to be slowed down by manual input can use voice recognition software to dramatically reduce document creation time.

It decreases work as all operations are done through voice recognition and hence paper work decreases to its maximum and the user can feel relaxed irrespective of the work.

- **Usability of other languages increases**.

As the speech recognition technology needs only voice and irrespective of the language in which it is delivered it is recorded, due to this perspective this is helpful to be used in any language.

- **Can help people with disabilities**

More recently student with learning or physical disabilities have been able to use SRS.Those with learning disabilities that affect their ability to write can now complete exame via voice

recognition technology and those with physical disabilities such as upper body paralysis can use SRS to communicate effectively with others.

- **Security**

With this technology a powerful interface between man and computer is created as the voice reorganization understands only the prerecorded voices and hence there are no ways of tampering data or breaking the codes if created.

- **Diminishes spelling mistakes**

Even the most experienced typists will occasionally have a spelling blunder, the a verage person is likely to make several mistakes in his or her composition.SRS always provides the ever the most experienced typists will occasionally have a spelling blunder,the average person is likely to Smake several mistakes in his or her composition.SRS always provides the correct splling of a word (assuming it translated it accurately in the first place),thus eliminating the need to spend time running spell checkers.

3.2 Disadvantages of automatic Speech recognition

- **Inaccuracy and slowness**

most people cannot type as fast as they speak. In theory this should make voice recognition software faster typing for entering text on a computer. However this may not always be the case because of the proofreading and correction required after dictating a document to the computer. Although voice recognition software may interpret your spoken words correctly the majority of the time ,you might still need to make corrections to punctuation. Additionally, the software may not recognize words such as brand names or uncommon surname until you add them to the programs library of words.SRS system are unable to recognize the words which are phonetically similar. E.g. there and their.

- **Adaptability**

Speech recognition software are not capable of adapting to various changing conditions which include different microphone, background noise, new speaker, new task domain new language even. The efficiency of the software degrades drastically.

- **Portability** – independence of computing platform
- **Out-of –vocabulary (oov) words**

Systems have to maintain a huge vocabulary of word of different language and sometimes according to the user phonetics also. They are not capable of adjust their vocabulary according to the change in users. System must have some method of detecting OOV words,and dealing with them in a sensible way.

- **Spontaneous speech**

System are unable to recognize the speech properly when it contain disfluencies(filled pauses, false starts,hesitations,ungrammatical construction etc). Spontaneous speech remain a problem.

4. How does the automatic Speech recognition technology work

When a person speaks, compressed air from the lungs is forced through the vocal tract as a sound wave that varies as per the variations in the lung pressure and the vocal tract. This acoustic wave is interpreted as speech when it falls upon a person's ear. In any machine that records or transmits human voice, the sound wave is converted into an electrical analogue signal using a microphone. When we speak into a telephone receiver, for instance, its microphone converts the acoustic wave into an electrical analogue signal that is transmitted through the telephone network. The electrical signals strength from the microphone varies in amplitude over time and is referred to as an analogue signal or an analogue waveform. If the signal results from speech, it is known as a speech waveform. Speech waveforms have the characteristic of being continuous in both time and amplitude.

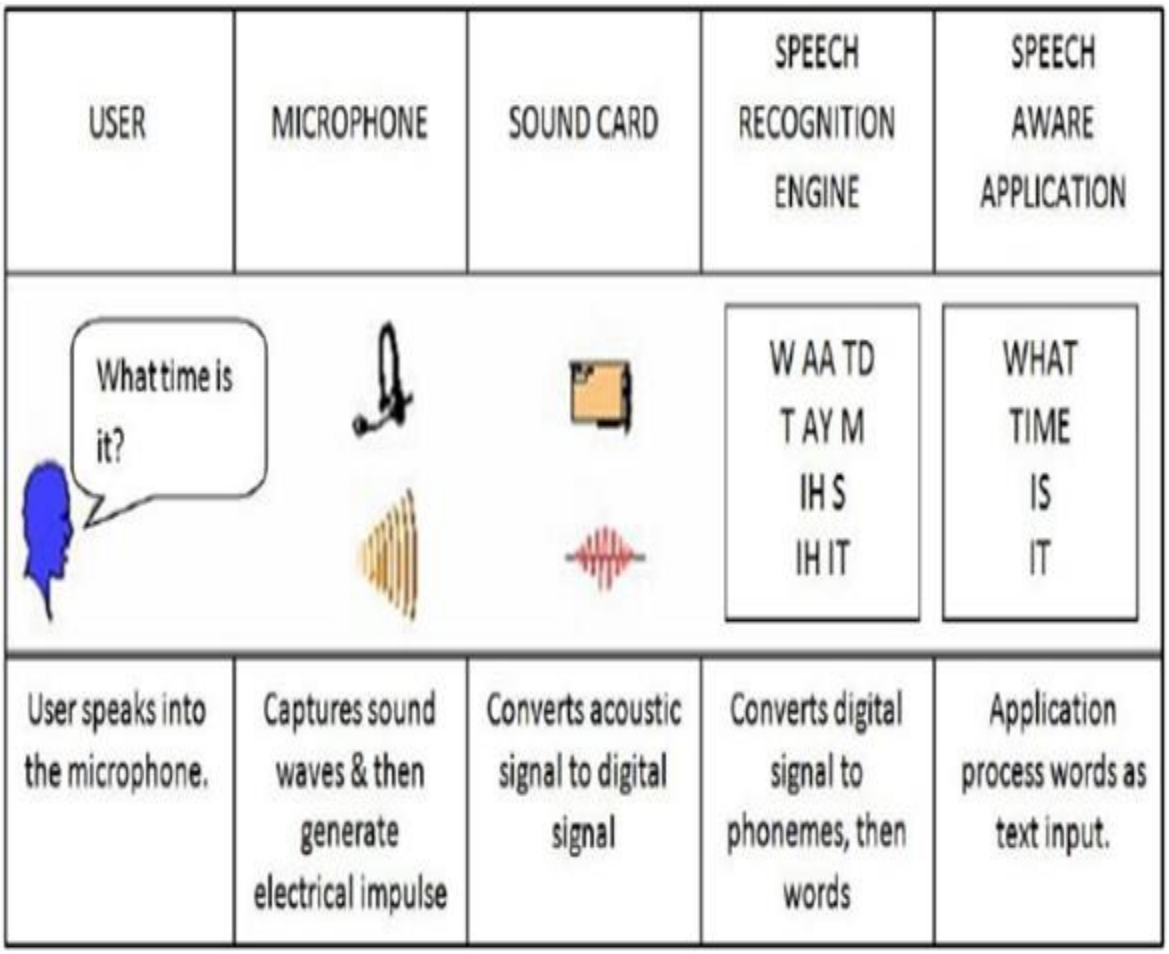

Figure1. 1Automatic Speech Recognition

5. Major Components of Automatic Speech Recognition

The major components of Automatic Speech Recognition are:-

Voice Input

With the help of microphone audio is input to the system, the pc sound card produces the equivalent digital representation of received audio.

Digitization

The process of converting the analog signal into a digital form is known as digitization. it involves the both sampling and quantization processes. Sampling is converting a continuous signal into discrete signal, while the process of approximating a continuous range of values is known as quantization.

Language Modeling

Language modeling is used in many natural language processing applications such as speech recognition tries to capture the properties of a language and to predict the next word in the speech sequence. The software language model compares the phonemes to words in its built in dictionary.

Acoustic Model

An acoustic model is created by taking audio recordings of speech, and their text transcriptions, and using software to create statistical representations of the sounds that make up each word. It is used by a speech recognition engine to recognize speech. The software acoustic model breaks the words into the phonemes.

Speech engine

The job of speech recognition engine is to convert the input audio into text to accomplish this it uses all sorts of data, software algorithms and statistics. Its first operation is digitization as discussed earlier, that is to convert it into a suitable format for further processing. Once audio signal is in proper format it then searches the best match for it. It does this by considering the words it knows, once the signal is recognized it returns its corresponding text string.

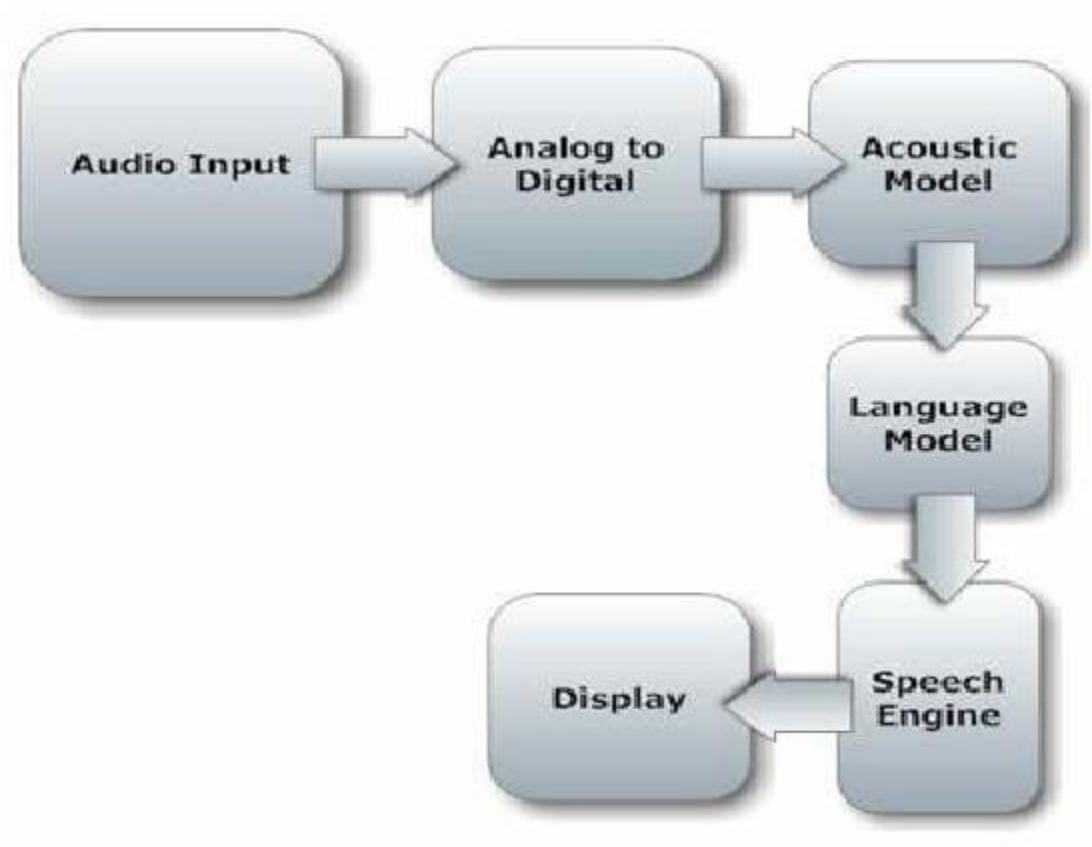

Figure1. 2 Major Components of Automatic Speech Recognition

6. major steps of automatic speech recognition

Any automatic speech recognition system involves following five major steps.

1. **Signal processing**

The sound is received through the microphone in the form of analog electrical signals noise is then removed and the signals are converted into digital signal. These digital signal are converted into a sequence of feature vectors.

Feature vector:- if you have a set of number representing certain features of an object you want to describe, it is useful for further processing to construct a vector out of these number by assigning each measured value to one component of the vector.

2. **Speech recognition**

This is the most important part of this process here the actual recognition is done. The sequence of feature vectors is then decoded into a sequence of word. This decoding is done on the basis of algorithms such as hidden markov model, neural network or dynamic time wrapping. The program has big dictionary of popular words that exist in language. Each feature vector is matched against the sound and converted into appropriate character group. It check and compares words that are similar in sound with the formed character group. All these similar word are then collected.

3. **Semantic interpretation**

Here it checks if the language allows a particular syllable to appear after another. A fter that ,will be grammar check.it tries to find out whether or not the combination of words any sense.

4. **Dialog management**

The errors encountered are tried to be corrected. Then the meaning of the combined word is extracted and the required task is performed.

5. **Response generation**

After the task is performed, the response or the result of that task is generated. The response is either in the form of a speech or text. What words to use so as to maximize the user understanding here. If the response is to be given in the form of speech, then text to speech conversion process is used.

7. Structure of standard speech recognition system

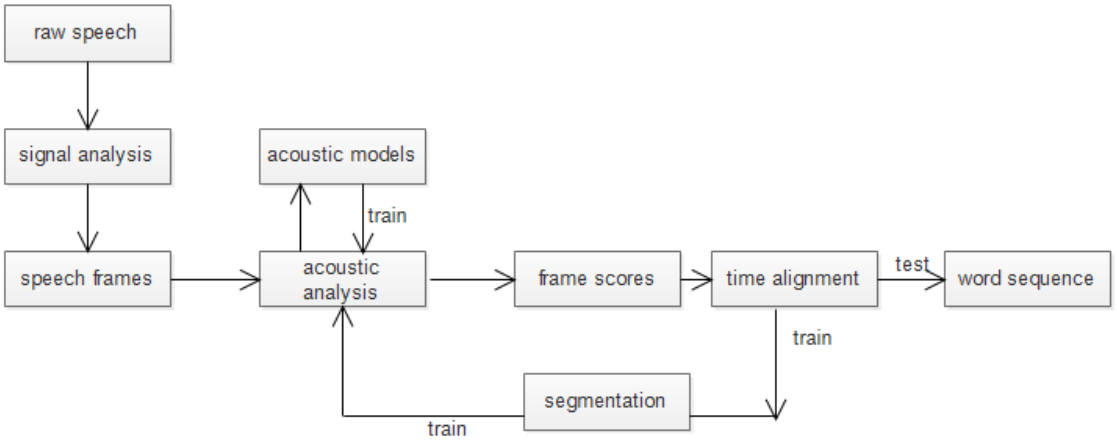

Figure1. 3 **typical speech recognition system**

The structure of a speech recognition system is illustrated in figure 1.3. The elements are as follows:

- **Raw speech**:-speech is typically sampled at a high frequency example,16 KHZ over a microphone or 8 KHZ over a telephone. This yields a sequence of amplitude values over time.
- **Signal analysis**:-raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information.

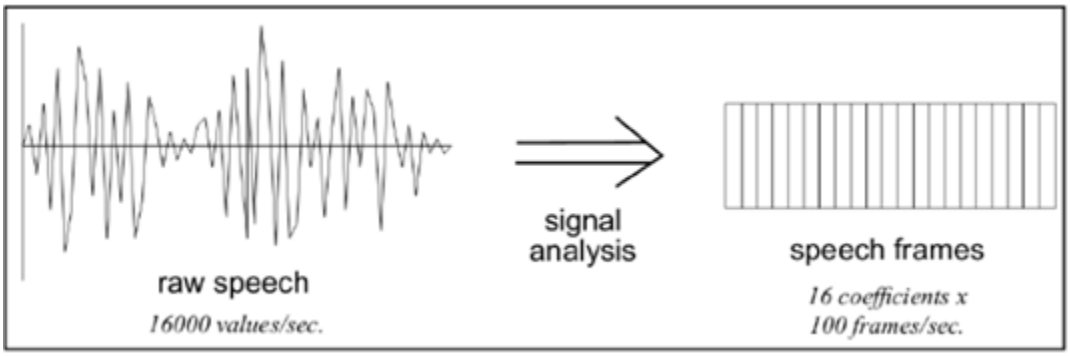

Figure1. 4 **signal analysis converts raw speech to speech frames**

- **Speech frame**: - the result of signal analysis is sequence of speech frames,typically at 10 milliseconds intervals,with about 16 coefficients per frame.these frames maybe augmented by their own first and second derivatives, providing. The speech frames are used for acoustic analysis.
- **Acoustic models**:-In order to analysis the speech frames for their acoustic content, we need a set of acoustic models. These are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties. During training, the acoustic models are incrementally modified in order to optimize the overall performance of the system. During testing, the acoustic models are left unchanged.

8. Type of Automatic speech recognition system

Automatic speech recognition system can be separated in several different classes by describing what type of utterances they have the ability to recognize. These classes are based on the fact that one of the difficulties of SR is the ability to determine when a speaker start and finishes an utterance. Most packages can fit in to more than one class, depending on which mode they are using.

8.1 Isolated word

Isolated words usually involve a pause between two utterances; it doesn't mean that it only accepts a single word but instead it requires one utterance at a time .

8.2.Connected word

Connected word system (or more correctly' connected utterances') are similar to isolated word, but allow separate utterances to be run together with aminimal pause them.

8.3. Continuous speech

Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries.continous speech recognizer allow users to speak almost naturally, while the computer determines content.

8.4. Spontaneous speech

At a basic leve,it can be thought of as speech that is natural sounding and not rehearsed. anasr system with spontaneous speech ability should be able to handle a variety of natural speech features such as word being run together ,ums"and ahs ,and even slight stutters.

8.5.Voice verification/identification

Some ASR system have the ability to identify specific users by characteristics of their voice. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity.

8.5.1 type of voice verification/identification

There are two types of voice verification/identification system,which are as follows:

Text-dependent:-if the text must be the same for enrollment and verification this is called Text-dependent .in a Text-dependent system, prompts can either be common across all speaker or knowledge-based information can be employed in order to create amulti-factor authentication scenario.

Text- independent:- Text- independent system are most often used for speaker identification as they require very little if any cooperation by the speaker.in this case the text during enrollment and test is different. As text-independent technologies do not compare what was said an enrollment and verification, verification applications tend to also employ speech recognition to determine what the user is.

9. Automatic Speech recognition algorithms

Dynamic time warping

Dynamic time warping algorithms is one of the oldest and most important algorithms in speech recognition. The simplest way to recognize an isolated word sample is to compare it against a number of stoted word templates and determine the best match. This goal depends upon a number of factors. First ,different samples of a given word will have somewhat different duration. This problem can be eliminated by simply normalizing the templates and the unknown

speech so that they all have an equal duration. However, another problem is that the rate of speech may not be constant throughout the word.

Hidden markov model

HMM can be used to model an unknown process that produces a sequence of observable outputs at discrete intervals, where the outputs are members of some finite alphabet. These models are called hidden Markov models precisely because the state sequence that produced the observable output is not known- its hidden.
HMM is represented by a set of states, vectors defining transitions between certain pairs of those states, probabilities that apply to state to state transitions, sets of probabilities characterizing observed output symbols, and initial conditions and the most flexible and successful approach to speech recognition so far has been hidden markov models(HMM). Hidden markov modelis a collection of sate connected by transition.it begin with a designated initial state .in each discrete time step, transition is taken up to a new state, and then one output symbol is generated in the state .the choice of transition and output symbol are both randam,governed by probability distribution.

An example is shown in the state diagram where states are denoted by nodes and transitions by directed arrows (vectors) between nodes. The underlying model is a markov chain. The circles represent states of the speaker's vocal system specific configuration of tongue, lips, etc that produce a given sound. The arrows represent possible transitions from one state to another. At any given time, the model is said to be in one state. At clock time, the model might change from its current state to any state to any state for which a transition vector exists. Transition may occur only from the tail to the head of a vector. A state can have more than one transition leaving it and more than one leading to it.

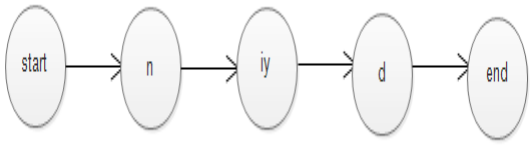

Figure1. 5 word model of need

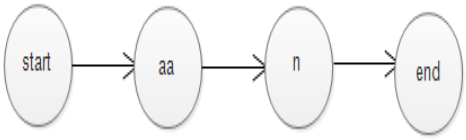

Figure1. 6 word model of on

Neural networks

A neural network consists of many simply processing units each of which is connected to many other unity. Each unit has a numerical activation level (analogous to the firing rate of real neurons).the only computation that an individual unit can do is to compute anew activation level based on the activation of the unit it is connected to. The connections between unit are weighted and the new activation is usually calculated as a function of the sum of the weighted inputs from other units.

10. Applications of automatic Speech recognition

From military perspective

Speech recognition programs are important from military perspective; in Air Force speech recognition has definite potential for reducing pilot workload. Beside the Air force such Programs can also be trained to be used in helicopters, battle management and other applications.

Document editing

This is a scenario in which one or both modes of speech recognition could be used to dramatically improve productivity. Dictation would allow users to dictate entire documents without typing. Command and control would allow user to modify formatting or change without using the mouse or keyboard.

For example a word processes might provide commands like:-bold ,italic change to time new roman font use bullet list text style and use 18 point type.

Speaker identification

Recognizing the patterns of speech of a various person can be used to identify them separately.it can be used as a biometric authentication system in which the user authenticates him/her with the help of their speech. The various characteristics of speech which involves frequency, amplitude and other special features are captured and compared with the previously stored data base.

Automation at call centers

Receiving call from a huge number of customers, answering them or diverting them to a particular customer care representative according to the customers demand.it can be used to provide a faster response to the customer and provide better service.

Medical disabilities

This technology is a great boon for blind and handicapped as they can utilize the speech recognition technology for various works. Those who are unable to operate the computer through keyboard and mouse can operate it with just their voice.

11. Conclusion and future enhancement

11.1 Conclusion

Speech recognition will revolutionize the way people interacted with smart devices and will ultimately differentiate the upcoming technologies. Almost all the smart devices coming today in the market are capable of recognizing speech. Many areas can benefit from this technology. Speech recognition can be used for intuitive operation of computer-based system in daily life.

This technology will spawn revolutionary changes in the modern world and become a pivot technology.

11.2. Future enhancement

Nowadays, the world is highly becoming a competitive world. Organizations have to divert their attention on using the recent technology to be on the first line and competitive. As described in our report this automatic speech recognition system has its own limitation. So there are some enhancements can be made in the future to make the system perfectly functional by solving its limitation. Some of the enhancements can be made are:-

- Microphone and sound systems will be designed to adapt more quickly to changing background noise levels, different environments, with better recognition of extraneous material to be discarded.
- Greater use will be made of "intelligent systems" which will attempt to guess what the speaker intended to say, rather than what was actually said, as people often mis-speak and make unintentional mistakes.
- Dictation speech recognition will gradually become accepted.

Reference

[1].JOE TEBELSKIS {1995}, SPEECH RECOGNITION USING NEURAL NETWORKS, School of Computer Science, Carnegie Mellon University.

[2].KÅRE SJÖLANDER {2003}, An HMM-based system for automatic segmentation and alignment of speech, Umea University, Department of OE TEBELSKIS {1995}, SPEECH RECOGNITION USING NEURAL NETWORKS, School of Computer Science, Carnegie Mellon University.

[3].KÅRE SJÖLANDER {2003}, An HMM-based system for automatic segmentation and alignment of speech, Umea University, Department of Philosophy and Linguistics.

[4] P. R. Dixon, T. Oonishi, and S. Furui. Fast acoustic computations using graphics processors. In Proc. IEEE Intl.Conf. on Acoustics, Speech, and Signal Processing (ICASSP),Taipei, Taiwan,2009.

[5] V. Gadde, A. Stolcke, D. Vergyri, J. Zheng, K. Sonmez, and A. Venkataraman. Building an ASR system for noisy environments: SRI's 2001 SPINE evaluation system.