INTERACTING WITH COMPUTER USING EARS AND TONGUE

Urmila Shrawankar[1] Anjali Mahajan[2]

[1]*Dept. of Information Technology,Government Polytechnic Institute,Nagpur– 440 001- INDIA*
urmilas@rediffmail.com Cell no. : +919422803996

[2]*Dept. of Computer Sci. & Engg G.H. Raisoni College of Engg., Nagpur– 440 016– INDIA*
armahajan@rediffmail.com

Abstract: Human computer interaction is concerned in the way Users (humans) interact with the computers. Some users can interact with the computer using the traditional methods of a keyboard and mouse as the main input devices and the monitor as the main output device. Due to one or another reason, some users are enable to interact with machines using a mouse and keyboard device, hence there is need for special devices. If we use computer for more time it is really difficult to sit on the chair, keeping hands continuously on the keyboard or mouse and keep watching the monitor.

To relax our body and interact comfortably with computer, we need some special device or method, so that computer will understand and accept commands without keyboard or by clicking mouse.

Speech Recognition System helps users who are unable to use traditional Input and Output (I/O) devices. Since four decades, man has been dreaming for an ”intelligent machine” which can master the natural speech. In its simplest form, this machine should consist of two subsystems, namely Automatic Speech Recognition (ASR) and Speech Understanding (SU). The goal of ASR is to transcribe natural speech while SU is to understand the meaning of the transcription. Recognising and understanding a spoken sentence is obviously a knowledge-intensive process, which must take into account all variable information about the speech communication process, from acoustics to semantics and pragmatics

Key words: Automatic Speech Recognition ,Text-To-Speech, Speech-To-Text, Interactive Voice Response–Systems, Linear Prediction Coding**,** Hidden Markov Model

1. INTRODUCTION

Speech is one of the oldest and most natural means of information exchange between human beings. We, as humans speak and listen to each other in human-human interface. The speech generation and production system is described in following figure.

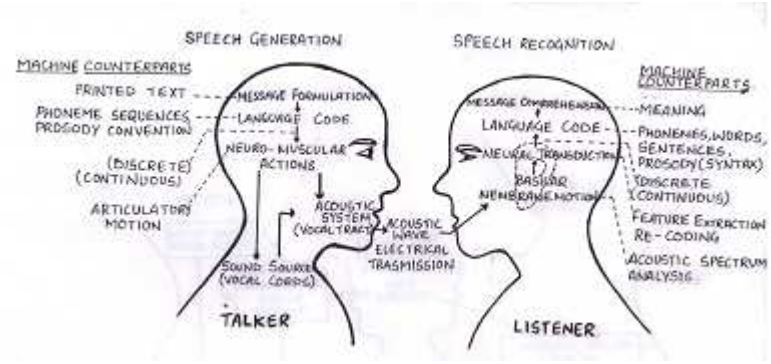


Fig.: Speech Generation and Production system.

Voice/speech recognition is a field of computer science that deals with designing computer systems that recognize spoken words. It is a technology that allows a computer to identify the words that a person speaks through a microphone or telephone.

Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone to a set of words.

Automatic Speech Recognition (ASR) is one of the fastest developing fields in the framework of speech science and engineering. In the new generation of computing technology, it comes as the next major innovation in man-machine interaction, after functionality of Text-To-Speech (TTS), supporting Interactive Voice Response (IVR) systems.

Nowadays, the statistical techniques prevail over ASR applications. Common speech recognition systems, these days can recognize thousands of words. The evolution of ASR, has improved its scope of applications in many aspects of daily life, for example, telephone applications, applications for the physically handicapped and illiterates and many others in the area of

computer science. Speech recognition is considered as an input as well as an output during the Human Computer Interaction (HCI) design. HCI involves the design implementation and evaluation of interactive systems in the context of the users' task and work

2. SPEECH RECOGNITION TECHNIQUES

Speech recognition techniques are as follows:

i. Template based approaches matching: Unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match. This has the advantage of using perfectly accurate word models. But it also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. Dynamic time warping is such a typical approach.

In this approach, the templates usually consist of representative sequences of feature vectors for corresponding words. The basic idea here is to align the utterance to each of the template words and then select the word or word sequence that contains the best. For each utterance, the distance between the template and the observed feature vectors are computed using some distance measure and these local distances are accumulated along each possible alignment path. The lowest scoring path then identifies the optimal alignment for a word and the word template obtaining the lowest overall score depicts the recognised word or sequence of words.

ii. Knowledge based approaches: An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully.

Thus this approach was judged to be impractical and automatic learning procedure was sought instead.

iii. Statistical based approaches: In this variation, speech is modeled statistically using automatic, statistical learning procedure, typically the Hidden Markov Models (HMM). The approach represents the current state of the art. The main disadvantage of statistical models is that they must take priori-modeling assumptions, which are liable to be inaccurate, handicapping the system performance. In recent years, a new approach of challenging problems of conversational speech recognition has emerged, holding a promise to overcome some fundamental

limitations of the conventional Hidden Markov Model (HMM) approach.

This new approach is a radical departure from the current HMM-based statistical modeling approaches. Rather than using a large number of unstructured Gaussian mixture components to account for the tremendous variations in the observable acoustic data of highly co-articulated spontaneous speech, the new speech model that have developed provides a rich structure for the partially observed (hidden) dynamics in the domain of vocal-tract resonance.

iv. Learning based approaches: To overcome the disadvantage of the HMMs, machine learning methods could be introduced such as neural networks and genetic algorithm / programming. In these machine-learning models, explicit rules or other domain expert knowledge need not be given. They can be learned automatically through emulations or evolutionary process.

v. The artificial intelligence approach attempts to mechanise the recognition procedure according to the way a person applies his intelligence in visualizing, analysing, and finally making a decision on the measured acoustic features. Expert system is used widely in this approach.

3. MATCHING TECHNIQUES

Speech-recognition engines match a detected word to a known word using one of the following techniques.

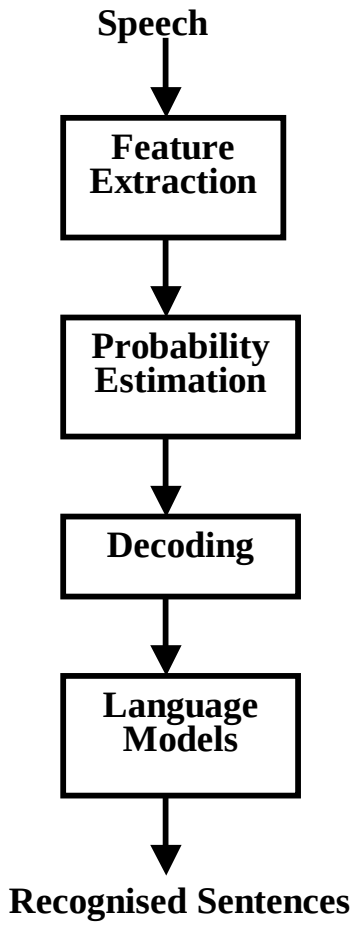
i. Whole-word matching: The engine compares the incoming digital-audio signal against a pre-recorded template of the word. This technique takes much less processing than sub-word matching, but it requires the user (or someone) pre-record every word that will be recognized (sometimes several hundred thousand words). Whole-word templates also require large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed.

ii. Sub-word matching: The engine looks for sub-words, usually phonemes and then performs further pattern recognition. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand. On discussing the research in the area of automatic speech recognition, it has been pursued for the last three decades, that only whole-word based speech recognition systems have found practical use

and became commercial successful. Though whole-word models became successful but the researchers mentioned that, they still suffer from two major problems, i.e. co-articulation problems and requiring a lot of training to build a good recognizer.

4. BUILDING AN APPLICATION BASED ON SPEECH INTERFACE

For building an application based on Speech Interface we need to follow following steps:



4.1 Input

Accept commands through Microphone.

4.2 Feature Extractions And Feature Matching

Feature extraction is the process that extracts a small amount of data from the voice that can later be used to represent each speech.
Feature matching involves the actual procedure to identify the unknown speech by comparing extracted features from his/her voice input.
All speech recognition systems have to serve two distinguished phases. The first one is referred to as enrollment sessions or training phase while the other is referred to as operation sessions or testing phase**.**

4.2.1 Speech Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representations for further analysis and processing. This is often referred to as the signal-processing front end.
A wide range of possibilities exists for parametrically representing the speech signal and the speech recognition task, such as Mel- Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC) and many more.

4.2.2 MFCC model :

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, that is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

4.2.3 LPC Model

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques and is a useful method for encoding quality speech at a low bit rate. It provides accurate estimates of speech parameters and efficient for computations. It is a speaker and text independent, normalized speech model and therefore LPC is more suitable for this type of application.

4.3 Training and Matching

In this part, by providing the Artificial Intelligence to the machine it is trained to match new sample of data with the trained samples.

4.3.1 Artificial Neural Networks

Neural Networks are often used as a powerful discriminating classifier for tasks in automatic speech recognition. They have several advantages over parametric classifiers. However, there are disadvantages in terms of amount of training data required and length of training time. The FeedForward BackPropagation model is commonly used for automatic speech recognition process.

4.3.2 Hidden Markov Model:

In the context of statistical methods for speech recognition, Hidden Markov Models (HMM) have become a well known and widely used statistical approach to characterize the spectral properties of frames of speech. It is a probabilistic model. As a stochastic modeling tool, HMMs have an advantage of providing a natural and highly reliable way of recognizing speech for a wide variety of applications. Since the HMM also integrates well into systems incorporating information about both acoustics and syntax, it is currently the predominant approach for speech recognition.

4.4 Text-to-Speech (Speech Synthesis)

After getting the results finally text results are converted into speech and outputted through speakers.

5. PROBLEMS IN DESIGNING SPEECH RECOGNITION SYSTEMS

ASR has been proved, that it is not an easy task. The main challenge in the implementation of ASR on desktops is the current existence of mature and efficient alternatives, the keyboard and mouse. In the past years, speech researchers have found several difficulties that contrast with the optimism of the first speech technology pioneers. According to Ray Reddy, in his review of speech recognition by machines says that the problems in designing ASR are due to the fact that it is related to so many other fields such as acoustics,

signal processing, pattern recognition, phonetics, linguistics, psychology, neuroscience, and computer science. And all these problems can be described according to the tasks to be performed.

i. Number of speakers: With more than one speaker, an ASR system must cope-up with the difficult problem of speech variability from one speaker to another. This is usually achieved through the use of large speech database as training data.
ii. Nature of the utterance: Isolated word recognition imposes on the speaker the need to insert artificial pause between successive utterances. Continuous speech recognition systems are able to cope-up with natural speech utterances in which words may be tied together and may at times be strongly affected by co-articulation. Spontaneous speech recognition systems allow the possibility of pause and false starts in the utterance, the use of words not found in the lexicon etc.
iii. Vocabulary size: In general, increasing the size of the vocabulary decreases the recognition scores.
iv. Differences between speakers due to sex, age, accent and so on.
v. Language complexity: The task of continuous speech recognizers is simplified by limiting the number of possible utterances through the imposition of syntactic and semantic constraints.
vi. Environment conditions: The sites for real applications often present adverse conditions (such as noise, distorted signal, and transmission line variability) that can drastically degrade the system performance.

6. CONCLUSION

The dream of a true virtual reality, a complete human-computer interaction system will not come true unless we try to give some perception to machine and make it perceive the outside world as humans do. Machine perception comes before any intelligent system consideration.

Speech understanding by the machine and interacting with the human like human-to-human will be the real interface for human-to-machine interaction.

The Speech interface will be a boon for physically challenged people, aged people and people having computer operation phobia.

Such applications can be further developed in different languages.

REFERENCES:

1. Bridle, J., Deng, L., Picone, J., Richards, H., Ma, J., Kamm, T., Schuster, M., Pike, S., Reagan, R., 1998. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, pp. 161.
2. Ma, J., Deng, L., 2004. Target-directed mixture linear dynamic models for spontaneous speech recognition. IEEE TRANSACTIONS ON SPEECH AND AUDIO ROCESSING, VOL. 12, NO. 1, JANUARY 2004.
3. Ma, J., Deng, L.,2004 A mixed-level switching dynamic system for continuous speech recognition. Elsevier Computer Speech and Language 18 (2004) 4965.
4. Mori R.D, Lam L., and Gilloux M. (1987). Learning and plan refinement in a knowledgebased system for automatic speech recognition. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 9(2):289-305.
5. Picheny, M., (2002). Large vocabulary speech recognition, *IEEE Computer*, 35(4):42-50.
6. Rabiner, L., R., and Wilpon, J. G., (1979). Considerations in applying clustering techniques to speaker-independent word recognition.*Journal of Acoustic Society of America.*66(3):663-673.
7. Reddy D.R., (1976). Speech Recognition by Machine: a Review. *Proceeding of IEEE,*64(4):501-531
8. Rudnicky, A.I., Lee, K.F., and Hauptmann, A.G. (1992) Survey of current speech technology. *Communications of the ACM,*37(3):52-57.
9. Svendsen T., Paliwal K. K., Harborg E., Husy P. O. (1989). Proc. ICASSP'89, Glasgow
10. Tolba, H., and O'Shaughnessy, D., (2001). Speech Recognition by Intelligent Machines, *IEEE Canadian Review* (38).
11. Wilpon J.G., D.M.DeMarco,R.P.Mikkilineni (1988) "Isolated word recognition over the DD telephone network -Results of two extensive field studies", Proc. ICASSP,pp. 55-58