IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. II (Jul.-Aug. 2016), PP 01-09
www.iosrjournals.org

Speech Recognition System – A Review

Shaikh Naziya S.[1*], R.R. Deshmukh[2]
[1*, 2]*Department of CSIT, Dr. B.A.M.University, Aurangabad, M.S., India*
naziyamsc@gmail.com

Abstract:*Language is the most important means of communication and speech is its main medium. In human to machine interface, speech signal is transformed into analog and digital wave form which can be understood by machine. Speech technologies are vastly used and has unlimited uses. These technologies enable machines to respond correctly and reliably to human voices, and provide useful and valuable services.The paper gives an overview of the speech recognition process, its basic model, and its application, approaches and also discuss comparative study of different approaches which are used for speech recognition system. The paper also provides an overview of different techniques of speech recognition system and also shows the summarization some of the well-known methods used in various stages of speech recognition system.*

Keywords: *Speech Recognition, Speech processing, Feature Extraction Techniques, Modeling Techniques, Applications of SRS.*

I. Introduction

Designing a machine that converse with human, particularly responding properly to spoken language has intrigued engineers and scientists for centuries. Speech Recognition System(SRS) is also known as Automatic Speech Recognition (ASR) or computer speech recognition which is the process of converting a speech signal to a sequence of words by means of an algorithm implemented as a computer program. It has the potential of being an important mode of interaction between humans and computers [1]. Today speech technology enabled applications are commercially available for a limited but interesting range of tasks. Very useful and valuable services are provided by these technology enabled machines, by responding correctly and reliably to human voices. In order to bring us closer to the "Holy Grail" of machines that recognize and understand fluently spoken speech, many important scientific and Technological advances have been took place, but still we are far from having a machine that mimics human behavior. Speech recognition technology has become a topic of great interest to general population, through many block buster movies of 1960's and 1970's [2].

The anthropomorphism of "HAL", a famous character in Stanley Kubrick's movie "2001: A Space Odyssey", made the general public aware of the potential of intelligent machines. In this movie, an intelligent computer named "HAL" spoke in a natural sounding voice and was able to recognize and understand fluently spoken speech, and respond accordingly. George Lucas, in the famous Star Wars saga, extended the abilities of intelligent machines by making them intelligent and mobile Droids like R2D2 and C3PO were able to speak naturally, recognize and understand fluent speech, move around and interact with their environment, with other droids, and with the human population. Apple Computers in the year of 1988, created a vision of speech technology and computers for the year 2011, titled "Knowledge Navigator", which defined the concepts of a Speech User Interface (SUI) and a Multimodal User Interface (MUI) along with the theme of intelligent voice-enabled agents. This video had a dramatic effect in the technical community and focused technology efforts, especially in the area of visual talking agents [2].

Languages, on which so far automatic speech recognition systems have been developed, are just a fraction of the total around 7300 languages. Chinese, English, Russian, Portuguese, Vietnamese, Japan, Spanish, Filipino, Arabic, Bengali, Tamil, Malayalam, Sinhala and Hindi are prominent among them [3].

A. Types of Speech Signal

In Speech Recognition System the ability to recognize the speech signal can be subdivided into different classes,

a. ***Isolated Words*****:** In this type, system accepts singleutterance at a time. And usually requires each utterance to have quiet on both side of sample window and require a speaker to wait between words. Its response will be better for single word but give poor result for multiple words input.

b. ***Connected Words:***In this type, multiple wordsgiven to the system which runs separately as isolated words and having small duration of time between them.

c. ***Continuous Speech:***In this type, natural speech isspoken by the user that is detectable by the machine. Continuous speech recognition is difficult to create because they utilize special method for implementation.

d. ***Spontaneous Speech:***In this type natural andspontaneous word has the ability to handle a variety of natural features such as words run together including mispronunciations, non-words and false statements, which are difficult to read.

B. Types of Speaker Model

Every speaker has unique properties which affects the voice. On the basis of these properties system is divided into two main classes.

a. ***Speaker Dependent Model:***Speaker dependentmodel depends on specific speaker. These models are easier to implement and less expensive. It gives more accurate result for specific speaker and less accurate result for other speakers.

b. ***Speaker Independent Model:*** Speaker independentmodels depend upon many speakers. These models are difficult to implement and more expensive. It gives more accurate result for many speakers and less accurate result for specific speaker.

C. Types of Vocabulary

The accuracy, complexity and processing requirements of the system depend upon the vocabulary size of speech recognition system. Some applications require a few words, others requires huge vocabulary. The types of vocabularies can be classified as follows.

a. Small size vocabulary includes tens of words.
b. Medium size vocabulary includes hundreds of words.
c. Large vocabulary size includes thousands of words.
d. Very large size vocabulary includes tens of thousands of words.
e. Out of size vocabulary includes mapping a word from the vocabulary into the unknown word.

D. Applications

The applications of speech recognition system (SRS) with their sector / area and fields are given in the table 1:

Table 1: Application of speech recognition systems

Sectors / Area / Field	Applications
Education sector	Speech to text processing, to correct pronunciation of vocabulary on foreign languages. Use keyboard to enter text verbally for handicapped students.
Medical Sector	Precision surgery, Automatic wheel chair, Medical transcription (digital speech to text)
Military Sector	Automatic aircraft control, helicopter, training air traffic controller, Automatic ammunition control.
Communication Sector	Voice dialing, telephone directory inquiry without operator assistance.
Domestic Sector	Ovens, refrigerators, washing machine, home appliances control etc.
General	Use of security purposes at highly secure places, Dictation system on market. To translate data from one language to another, video gaming and ATM (data entry).

II. General Steps For Speech Recognition Systems (Srs)

In speech recognition system an unknown speech signal is transformed into sequence of feature vectors by different speech processing techniques. It converts feature vector to phoneme lattice by applying an algorithms [4]. A recognition module transforms the phoneme lattice into a word lattice by lexicon and then grammar is applied to word lattice to recognize the specific words or text. Figure 1 shows the information for general steps in speech recognition system (SRS).

The speech recognition process is divided into several steps.

A. ***Step 1:***In this step speech signal is divided intoequally spaced blocks to get signal characteristics such as, total energy, zero crossing strength across various frequency ranges etc. By using these characteristics feature vectors combine each block with the phoneme to produce a string of phonemes.

B. ***Step 2:***In this step spectrum analysis is appliedon each block by using linear predictive coding technique, fast Fourier transform(FFT) and bank of frequency filters.

C. ***Step 3:***In this step decision process is performedon each block. Each phoneme has distinguished features which narrow the field.

D. ***Step 4:*** This step is used to enhance theperformance of decision process to get high degree of success using different algorithms. For each word of vocabulary an algorithm is constructed and then string of phonemes is compared against each algorithm.

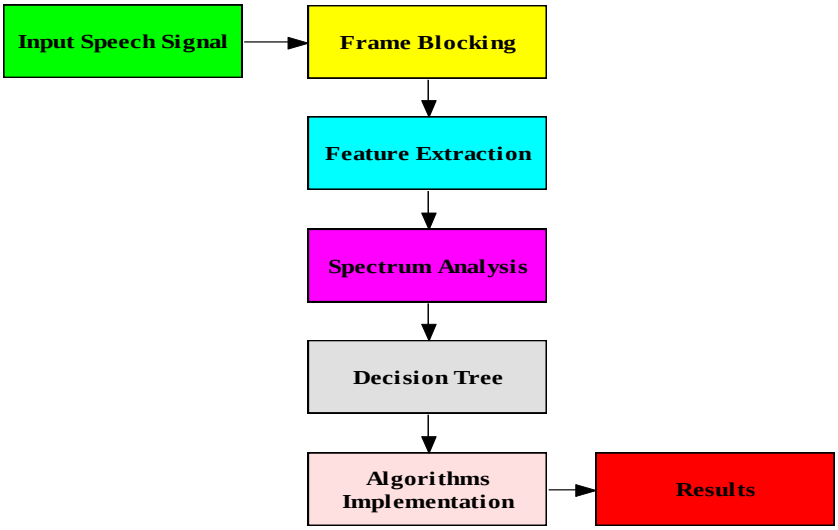

Figure 1: General steps for Speech Recognition System (SRS)

III. Techniques In Speech Recognition Systems

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories. Davis, Biddulph and Balashek developed an isolated digit recognition system for a single speaker. The speaker recognition system may be viewed as working in a four stages.

A. Speech Analysis
B. Speech Feature Extraction Techniques
C. Modeling
D. Testing/Matching techniques

A. Speech Analysis

In speech analysis technique Speech data contains different types of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The physical structure and dimension of vocal tract as well as excitation source are unique for each speaker. The speech analysis deals with stages with suitable frame size for segmenting speech signal for further analysis and extracting [5]. The speech analysis is done with following three techniques.

a. ***Segmentation Analysis:***In this case, speech is analyzed using the frame size and shift in the range of 10-30 ms toextract speaker information. Studies have been made in using segmented analysis to extract vocal tract information of speaker recognition.

b. ***Sub-segmental Analysis:***Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmentalanalysis. This technique is used mainly to analyze and extract the characteristic of the excitation state. [6]. The excitation source information is relatively fast varying compared to vocal tract information, so small frame size and shift are required to best capture the speaker-specific information [7].

c. ***Supra-segmental Analysis:*** In this case, speech is analyzed by using the frame size and shift of 100-300 ms to extractspeaker information mainly due to behavioral tract and here speech is analyzed using the frame size. This technique is used mainly to analyze and characteristic due to behavior character of the speaker. These include word duration, intonation, speaker rate, accent etc.

B. Speech Feature Extraction Techniques

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. The utterance can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

- Easy to measure extracted speech features
- It should not be susceptible to mimicry
- It should show little fluctuation from one speaking environment to another
- It should be stable over time
- It should occur frequently and naturally in speech

The most widely used feature extraction techniques are explained below,

a. Linear Predictive Coding (LPC)

One of the most powerful signal analysis techniques is the method of linear prediction. LPC [8] [9] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech [10]. The predictor coefficients are therefore transformed to a more robust set of parameters known as Cepstral coefficients. Figure 2 shows the steps involved in LPC feature extraction.

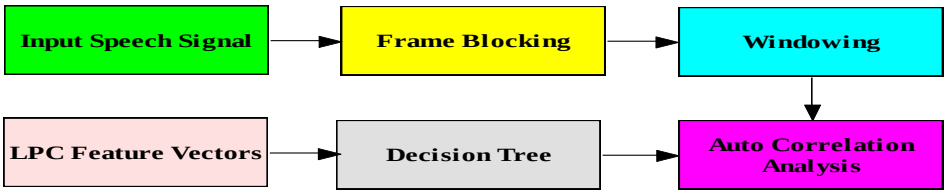

Figure 2: Steps involved in LPC Feature Extraction

b. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [8] [9] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [11], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the Cepstral coefficients calculation [8] [9]. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula.

$$Mel_f = 2595x \log_{10} (1 + F/700) \quad \text{… (Eq. 1)}$$

The figure 3 shows the steps involved in MFCC feature extraction.

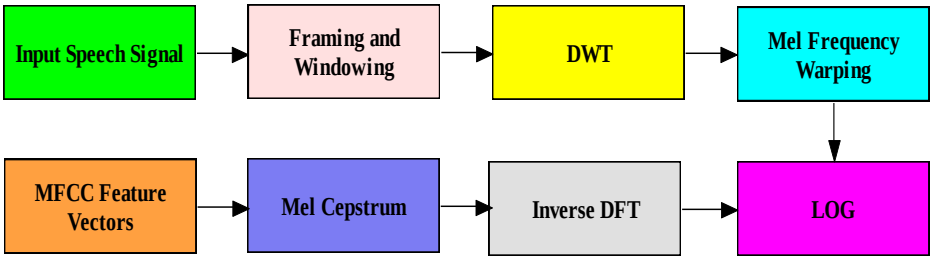

Figure 3: Steps involved in MFCC Feature Extraction

c. Modeling Technique

Speech recognition is a vast and emerging field nowadays. Every speech signal has different characteristics depend on utterances, to achieve this task of recognition different approaches are used. Each technique has its own benefits depends upon the scenario. Some approaches include template based approach, neural network based approach, statistical based approach, Hidden Markov Model (HMM) based speech recognition etc. Most famous of all is the HMM approach because It is easy, simple and reliable, it can be automatically trained and feasible to use. Some approaches are listed below.

i. Acoustic Phonetic Recognition

Acoustic phonetic recognition performs the function at phoneme level. It exist distinctive, finite phonemes which are characterized by a set of acoustic properties that occur in a speech signal [12]. English language includes forty different phonemes and doesn't depend on the vocabulary. It is the earliest approach of SRS to recognize speech by providing labels to the speech. This approach includes highly variable phonetic units, the variability in these unit are straight forward which are easily learned by machine [13]. Figure 4 shows

the information about Acoustic phonetic recognition block diagram.
This approach is divided into three steps

- Feature Extraction
- Segmentation and Labeling
- Word-level recognition

In first step the spectral analysis of speech signal along with feature detection are performed which convert spectral measurements to set of features that provide vast acoustic properties to the signal.

In second step attachment of phonetic label is done with segmentation region of speech signal. It gives phoneme lattice characterization of speech signal.

In the last step by using segmentation and labeling string of words are produced from phonetic label sequence.

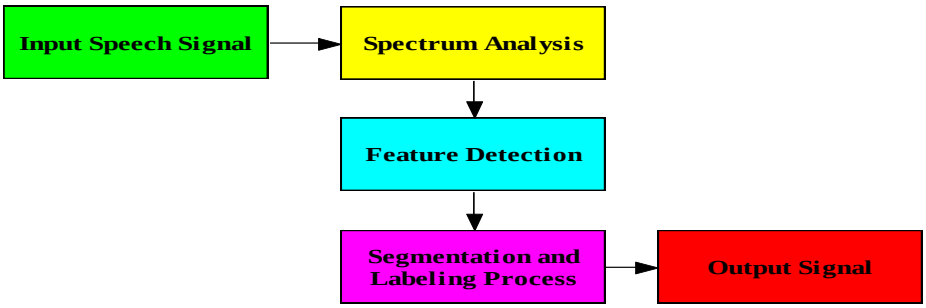

Figure 4: Acoustic Phonetic Recognition Block Diagram

ii. Pattern Recognition Approach

By using mathematical framework this approach formulates speech pattern representation from formal training algorithms by set of labeled training samples via formal training algorithms [14].
This approach involves two major steps.

- Pattern training
- Pattern compression

In pattern training process speech signals are shown in speech template or statistical modal.In pattern comparison process unknown speeches are compared with possible learned pattern which are formed from training stage. Figure 5 shows the pattern recognition approach block diagram.

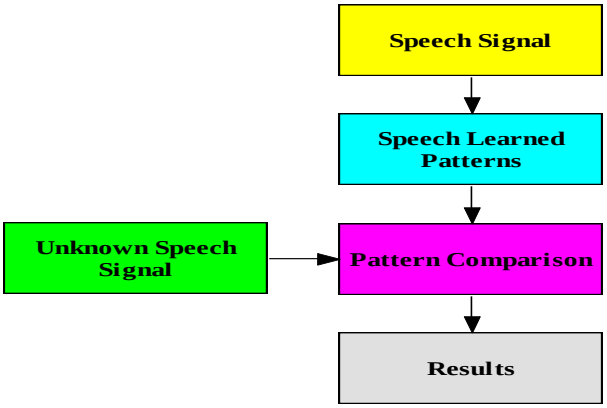

Figure 5:Pattern Recognition Approach Block Diagram

iii. Template Based Approach

In this approach by using a collection of speech pattern, dictionary of words are created which is stored as a reference, after that matching of unknown speech is done with the reference template [15]. Then selection of the best matching template is done. It is also a modified form of pattern recognition.It is the oldest and least effective method but dominating approach in 1950's to 1960's. It is still successful for small dictionaries and specifically for isolated word recognition. It is the primary technology for verification of speaker. Figure 6 shows the template based approach block diagram.

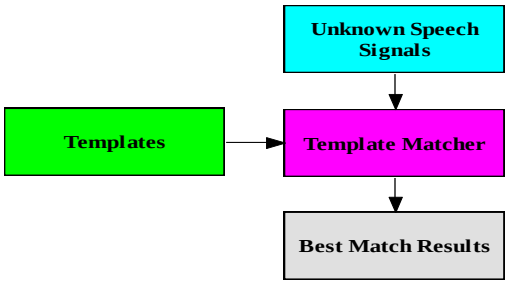

Figure 6:Template Based Approach Block Diagram

iv. Vector Quantization Approach

For efficient reduction of data of automatic SRS this approach is used, because in automatic SRS transmission rate is not a major issue. The efficiency of this approach lies on using best code book which yields the lowest distance measure [16]. To measure the average distance across the training frames and frame corresponding to longer segments which are more frequently used, the code book entries are selected for this purpose. To specify the code of words for less frequent frames such frames are more likely to be used especially smaller code books. Figure 7 shows the vector quantization approach block diagram.

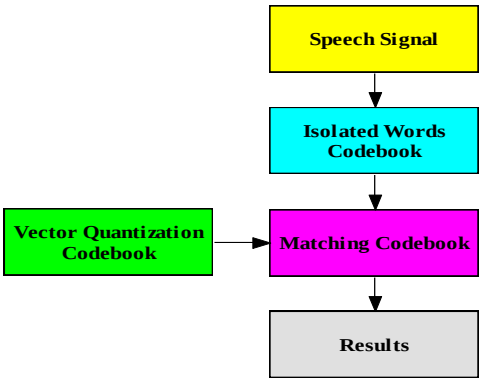

Figure 7:Vector Quantization Approach Block Diagram

v. Dynamic Time Wrapping

This approach finds similarities between two different sequences which vary with time and speech [17].Consider an example in which two cars are traveling on a same way, if one car runs faster than the other. Thus measuring of two sequences with respect to time is done. If both cars are accelerated and de-accelerated during the observation then it measures the sequences with respect to speed. This approach is called as wrapping due to its nonlinear variation. In DTW any type of machine matching is done with certain restrictions for isolated words, this approach is beneficial for connected words and it can be easily modified. If the sequence is independent of some nonlinear variation with respect to time then it measures the similarities in both of them. As the name suggests the optimization process is achieved by using dynamic programming. Figure 8 shows the dynamic time wrapping block diagram.

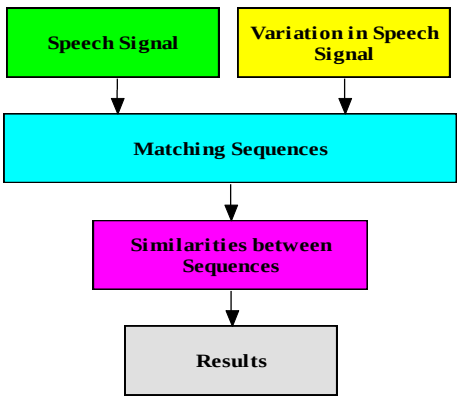

Figure 8:Dynamic Time Wrapping Block Diagram

vi. Statistical Based Approach

Uncertainty or incompleteness occurs in speech recognition due to many reasons like mixing of sound, speaker variability, confusable words etc. The word stochastic points towards the sequence of non-deterministic selections from set of alternatives. Non deterministic means choices depend on the characteristics of input. This technique uses best probabilistic model so it is most efficient technique for speech recognition. Hidden Markov modeling is the most popular technique of all characteristics of input. This technique uses best probabilistic model so it is most efficient technique for speech recognition [18]. Hidden Markov modeling is the most popular technique of all.

- ***Hidden Markov Model***

It is widely use stochastic approach today because in this approach speech is generating from number of states for each HMM model. In HMM we mixture multi vibrate Gaussian distribution, probabilistic mean, variance and mixture weight for speech [19]. Each phoneme has different output distribution. A HMM for a sequence of words or phoneme is made by concentrating the individual train HMM for the separate words and phoneme. Figure 9 shows the information about the hidden markov model (HMM)

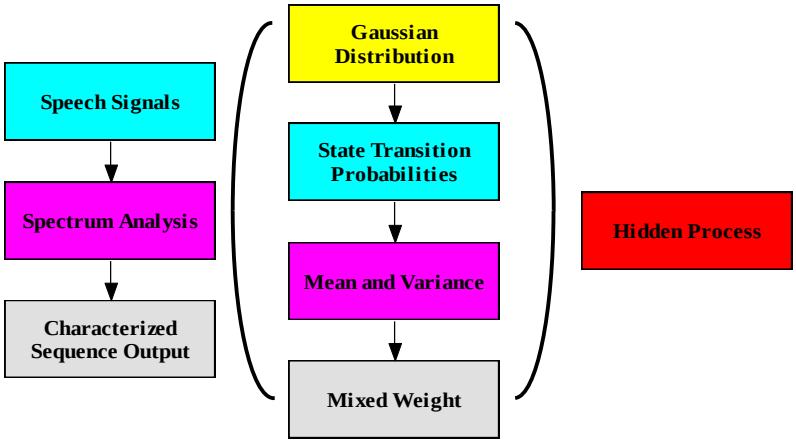

Figure 9: Hidden Markov Model(HMM) Block Diagram

vii. Artificial Neural Network Approach

This approach is designed for complicated tasks but it is not as efficient as HMM in the case of large vocabularies. Phoneme recognition is the general approach of neural networks. In this approach the technique of intelligence, analyzing and visualizing of speech signal is done to measure phonetic features [20].The network includes a huge number of neurons. Each neuron computers nonlinear weight of inputs and broadcast result to the outgoing units, training sets are used for assigning pattern of values to input and output neurons, training set determines the weight of strength of each pattern. Figure 10 shows the artificial neural network block diagram.

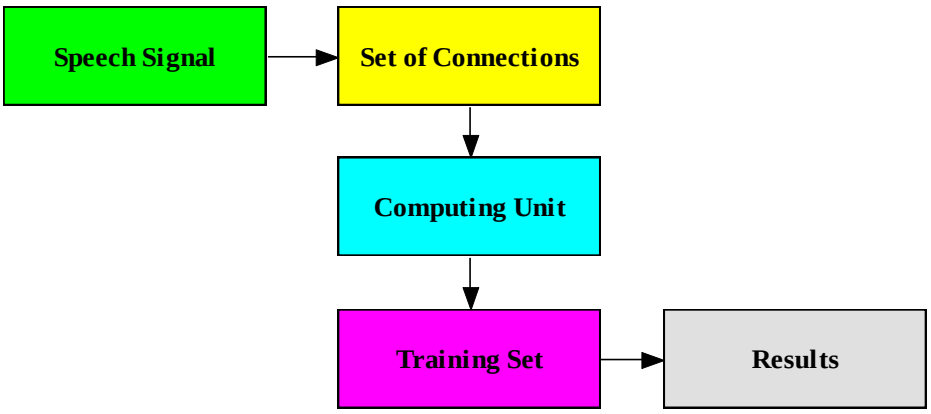

Figure 10: Artificial Neural Network Approach Block Diagram

d. Matching Techniques

Speech-recognition engines match a detected word to a known word using one of the following techniques (Svendsen et al., 1989)[21].

i. Whole-word matching

The engine compares the incoming digital-audio signal against a prerecorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord 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 [21].

ii. Sub-word matching

The engine looks for sub words usually phonemes and then performs further pattern recognition on those. 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 theword can be guessed from English text without requiring the user to speak the word beforehand. Discuss that research in the area of automatic speech recognition had been pursued for the last three decades [21].

IV. Comparative Study Of Speech Recognition System Approaches

The comparative study of speech recognition system approaches have given in the table 2 with their advantages and disadvantages:

Table 2: Comparative Study of Speech Recognition Modeling Technique

Sr. No.	SRS Techniques	Advantages	Disadvantages
1.	Acoustic Phonetic Recognition	1. It reduces Processing Time for connected words	1. Not widely used in commercial application due to large time execution of each isolated word
2.	Dynamic Time wrapping	1. Continuity is less important because it can match sequence with missing information. 2. Reliable time alignment between reference and test pattern.	1. It matches between two given sequence with certain restrictions. 2. It requires maximum time for complex computational work. 3. Limited number of templates.
3.	Pattern Recognition Approach	1. It recognize pattern quickly, easily and automatically because word to word matching will occur.	1. It is useful for word to word matching. 2. Template is the main problem. 3. Slow process. 4. It doesn't recognize speech if new variation of pattern occur.
4.	Vector Quantization Approach	1. It is useful for efficient data reduction.	1. It is text depended because need codebook for matching.
5.	Template Base Approach	1.It is better for discrete words 2. Less error occur due to segmentation and classification of small variable units.	1. Expensive due to large vocabulary size in each word has reference templates for it. 2. Template matching and preparation requires more time. 3. It difficult to recognize similar templates.
6.	Artificial Neural Network Approach	1. It can solve complex computational task effectively within less time. 2. It has ability to automatically train the data and taught the system changing from initial training model without error. 3. It can handle noisy, low quality data efficiently and require minimum training data vocabulary.	1. It gives inefficient result for large vocabulary 2. It is expensive because for training it requires much iteration over large amount of training data. 3. Full nature of neural network is not fully understood still. 4. It requires more training time 5. More error variation occurs due to complex architecture of neural networks.
7.	Statistical Based Approach(Hidden markov method)	1. The vocabulary size of HMM is very high so it can train large amount of data. 2. It has accurate mathematical frame work. 3. The trained algorithms are easily available. 4. It can implement easily and anyone can easily change the size, type and architecture of these models to suit particular word. 5. It is more robust because probability of certain words can occur next to each other. 6. It has capability to achieve recognition rates accurately. 7. It has efficient learning algorithm. 8. It has flexible and general model for sequence properties. 9. It can learn variable data unsupervisely.	1. A significant increase in computational complexity. 2. Need large amount of data.

V. Conclusion

Speech Recognition System (SRS) is growing day by day and has unlimited applications. Thestudy has shown the overview of the speech recognition process, its basic model, and applications. In this studytotal seven different approaches which are widely used for SRS have been discussed and after comparative study of these approaches it is concluded that Hidden markov method(HMM) is best suitable approach for a SRS because it efficient, robust, and reduces time and complexity.

References

[1]. Abdulla, W. H., Chow, D., & Sin, G. (2003, October). Cross-words reference template for DTW-based speech recognition systems. In *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region* (Vol. 4, pp. 1576-1579). IEEE.

[2]. Al-Qatab, B. A., & Ainon, R. N. (2010, June). Arabic speech recognition using hidden Markov model toolkit (HTK). In *Information Technology (ITSim), 2010 International Symposium in* (Vol. 2, pp. 557-562). IEEE.

[3]. Bahl, L. R., Brown, P. F., De Souza, P. V., & Mercer, R. L. (1993). Estimating hidden Markov model parameters so as to maximize speech recognition accuracy. *Speech and Audio Processing, IEEE Transactions on*,*1*(1), 77-83.

[4]. Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., & Picheny, M. A. (1993). A method for the construction of acoustic Markov models for words. *Speech and Audio Processing, IEEE Transactions on*, *1*(4), 443-452.

[5]. Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992, February). Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the workshop on Speech and Natural Language* (pp. 339-343). Association for Computational Linguistics.

[6]. Charles, A. H., & Devaraj, G. (2004). Alaigal-A Tamil Speech Recognition. *Tamil Internet*.

[7]. Dumitru, C. O., & Gavat, I. (2006, June). A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language. In *Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006 focused on* (pp. 115-118). IEEE.

[8]. Furui, S., Ichiba, T., Shinozaki, T., Whittaker, E. W., & Iwano, K. (2005). Cluster-based modeling for ubiquitous speech recognition. *Interspeech2005*, 2865-2868.

[9]. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, *10*(3), 16-24.

[10]. Ghai, W., & Singh, N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications*, *41*(8).

[11]. GIN-DER WU AND YING LEI (2008), "A register array based low power FFT processor for speech recognition. *"JOURNAL OF INFORMATION SCIENCE AND ENGINEERING* 3 (2008).

[12]. Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–A brief history of the technology development. *Encyclopedia of Language and Linguistics*.

[13]. King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, *121*(2), 723-742.

[14]. Klevans, R. L., & Rodman, R. D. (1997). *Voice recognition*. Artech House, Inc.

[15]. Maheswari, N. U., Kabilan, A. P., & Venkatesh, R. (2010). A hybrid model of neural network approach for speaker independent word recognition. *International Journal of Computer Theory and Engineering*, *2*(6), 912.

[16]. Moore, R. K. (1994, September). Twenty things we still don't know about speech. In *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research and Technology*.

[17]. Morales, N., Hansen, J. H., & Toledano, D. T. (2005, March). MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech. In *ICASSP (1)* (pp. 521-524).

[18]. Reddy, D. R. (1966). Approach to computer speech recognition by direct analysis of the speech wave. *The Journal of the Acoustical Society of America*, *40*(5), 1273-1273.

[19]. Shaughnessy, D. O. (2003). Interacting with computers by voice: automatic speech recognition and synthesis. *Proceedings of the IEEE*, *91*(9), 1272-1305.

[20]. Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., & Bell, D. (1989, May). Linguistic constraints in hidden Markov model based speech recognition. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on* (pp. 699-702). IEEE.

[21]. Yegnanarayana, B., Prasanna, S. M., Zachariah, J. M., & Gupta, C. S. (2005). Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. *Speech and Audio Processing, IEEE Transactions on*, *13*(4), 575-582.