Speech Recognition: A Review

Manoj Kumar Sharma
manoj186@yahoo.co.in
Omendri kumari
omifauzdar@gmail.com
School of Engineering & Technology, Jaipur National University – Jaipur

----------ABSTRACT----------

***Abstract* - Speech is a natural mode to interact with others. With speech, we can express our words to others. Speech recognition is a way or technology where the statements or commands of human speech to understand and react accordingly. Speech recognition allows machining system to turn the incoming speech signals into commands through the process of identifying and understanding. It also creates the natural voice communication function. Main Goal of speech recognition is to achieve better language communication between man and machine. So it is a great technology of human machine interface. The paper describes the speech recognition technology development is all basic principles, methods and classification of this technology. Accuracy of different methods of speech technology is provided to sort out methods with their performance aspect.**

***Keywords*: Speech; Speech recognition; voice; machine control; human machine interaction; communication; device control.**

---

1. Introduction

Speech is a natural way of communication for people, but sometimes it does not work i.e. disabled person. Over the last one decade, here is a need to enable human to communicate with machines without performing any text input. Here speech recognition is a technology can make capable disabled humans to communicate with machines with their disabilities. Speech recognition is a system that used by the human to listen, identify and understand what does the user want by speaking. It is a conversion of speech to text in a system.

Speech recognition is the machine on the statement or command of human speech to identify and understand and react accordingly. It is based on the voice as the research object, it allows the machine to automatically identify and understand human spoken language through speech signal processing and pattern recognition. The speech recognition technology is the high-tech that allows the machine to turn the voice signal into the appropriate text or command through the process of identifying and understanding. Speech recognition is a cross-disciplinary and involves a wide range. It has a very close relationship with acoustics, phonetics, linguistics, information theory, pattern recognition theory and neurobiology disciplines. With the rapid development of computer hardware and software and information technology, speech recognition technology is gradually becoming a key technology in the computer information processing technology [1]. The goal in automatic speech recognition is to provide a means for verbal human-to-machine communication. Although both speech coding and recognition involve analysis of the speech wave, the voice recognition problem is by far more difficult. Applications of speech recognition technologies include process automation, telephone inquiry, automatic banking, secure voice access, to name just a few. Although the research in speech recognition is allocation driven. There is a significant lack between research and commercial deployment [2]. Understanding speech requires the integration of a number of different and complex processes, such as, signal processing (recognition of phonemes, syllables and words), syntactic parsing and semantic analysis. Language is a system that enables a speaker to make more effective use of words that are usually learned during childhood. The characteristics of speech sounds depend on the particular human language or dialect [3]. Basically speech recognition is a pattern recognition problem. Speech Recognition Systems are generally classified as discrete or continuous systems that are speaker dependent, independent or adaptive.

A speaker-dependent system requires that the user record an example of the word, sentence or phrase prior to its being recognized by the system i.e. the user trains the system. Some speaker-dependent systems require only that the user record a subset of system vocabulary to make the entire vocabulary recognizable. A speaker-independent system does not require any recording prior to system use. It is developed to operate for any speaker of a particular type. A speaker adaptive system is developed to adapt its operation to the characteristics of new speaker [4].

1.1 Types of Speech :

The hierarchy of speech recognition problems which have been attacked, and the resulting application tasks which became viable as a result, includes the following [6]

TABLE 1: TYPES of SPEECH

Type of speech	Specification
Isolated Word	Listen –non listen state
Connected Word	Run together minimum pause
Continuous Speech	Natural continues speech
Spontaneous Speech	Variety of natural speech

1.1.1 Isolated Word

Isolated word recognizes attain usually require each utterance to have quiet on both sides of sample windows. It accepts single words or single utterances at a time .This is having "Listen and Non Listen state". Isolated utterance might be a better name of this class [5]. Isolated word recognition-both speakers trained and speaker independent. This technology opened up a class of applications called 'command and control' applications in which the system was capable of recognizing a single word command (from a small vocabulary of single word commands), and appropriately responding to the recognized command. One key problem with tliis technology was the sensitivity to background noise (which were often recognized as spurious spoken words) and extraneous speech which was inadvertently spoken along with the command word. Various types of 'keyword spotting' algorithms evolved to solve these types of problems [6].

1.1.2 Connected Word

The Connected word system is similar to isolated words, but allow the separate utterance to be "run together minimum pause between them [5]. Connected word recognition-both speakers trained and speaker independent. This technology was built on top of word recognition technology, choosing to exploit the word models that were successful in isolated word recognition, and extend the model to recognize a concatenated sequence (a string) of such word models as a word string. This technology opened up a class of applications based on recognizing digit strings and alphanumeric strings, and led to a variety of systems for voice dialing, credit card authorization, directory assistance lookups, and catalog ordering [6].

1.1.3 Continuous speech

Continuous speech recognizers allows user to speak almost Naturally, while the computer determine the contents. Recognizer with continued speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries [5]. Continuous or fluent speech recognition-both speakers trained and speaker independent. This technology led to the first large vocabulary recognition systems which were used to access databases (the DARPA Resource Management Task), to do constrained dialogue access to information (the DARPA ATIS Task), to handle very large vocabulary read the speech for dictation (the DARPA NAB Task), and eventually were used for desktop dictation systems for PC environments [6].

1.1.4 Spontaneous speech

At a basic level, it can be thought of as speech that is natural Sounding and not rehearsed. An ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together [5]. Spontaneous conversation systems which are able to both recognize the spoken material accurately and understand the meaning of the spoken material. Such systems, which are currently beyond the limits of the existing technology, will enable new services such as 'Conversation Summarization', 'Business Meeting Notes', 'Topic Spotting' in fluent speech (e.g., from radio or TV broadcasts), and ultimately even language translation services between any pair of existing languages [6].

Speech recognition uses several techniques to "recognize" the human voice. It functions as a pipeline that converts digital audio signals coming from the sound card to recognize speech. These signals pass through several stages, where various mathematical and statistical methods are applied to figure out what is actually being said [4].

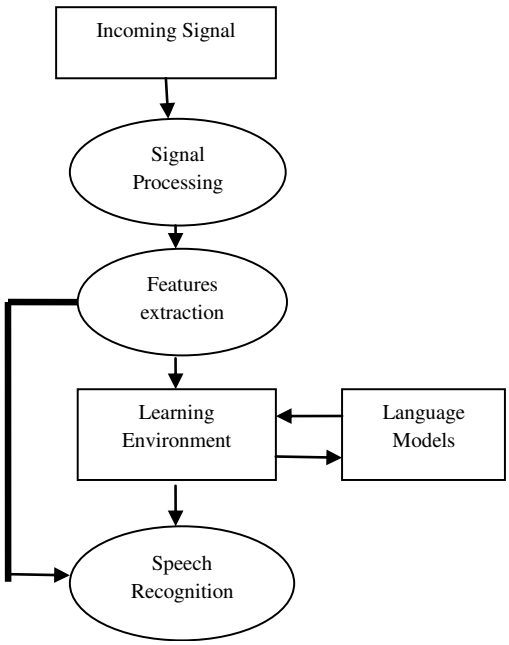

Figure 1: verification in the speech recognition process

Figure 1 shows the four steps to perform identification or verification in the speech recognition process.

First perform, convert the speech into digital wave. A simple way to convert the sound wave into digital wave via microphone. It's called signal processing. Second, compute features. In this feature are computed. Every 10 milliseconds, with 10 ms section called a frame and its process called feature extraction. Third, a neural network. Here we pre construct a grammar to for neural learning, here learning is performed by language models. Fourth, match the words. Speech recognition output that is performed on words.

2. Basic Principle Of Speech Recognition

The speech recognition system is essentially a pattern recognition system, including feature extraction, pattern matching and the reference model library [1].

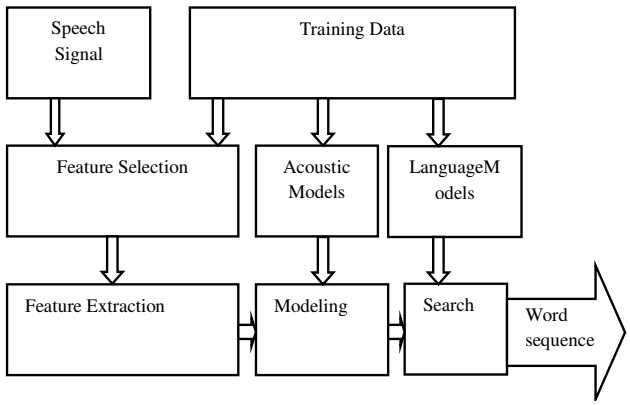

Figure 2: Principle of Speech Recognition

The unknown voice through the microphone is transformed into an electrical signal on the input of the identification system, the first after the pretreatment. The system establishes a voice model according to the human voice characteristics, analyzes the input voice signal and extracts the required features on this basis, it establishes the required template of the speech recognition. The computer is used in the recognition process according to the model of the speech recognition to compare the voice template stored in the computer and the characteristics of the input voice signal. Search and matching strategies to identify the optimal range of the input voice matches the template. According to the definition of this template through the lookup table can be given the recognition results of the computer [1]

A) Speech Signal:
First, the speech signal is treated graphically. The signal image requires some preprocessing steps to be followed spectral estimation method based on the linear predictive coding principle. This model is built on the idea of prediction error minimization. The signal spectrum obtained forms the basis for the analysis of Toeplitz Matrix (TM). Then TM is applied to describe the speech signal in the form of a feature vector for each signal image. Afterwards, the Toeplitz-based feature vector enters the classification stage. For classification, probabilistic and radial basis functions (RBF) NNs have been used [4].

B) Feature Selection:
There are several choices of features for speech recognition, i.e. amplitude, zero-crossing rate, and spectral content. Amplitude (or power) is a primary source for endpoint information and also for vowel/consonant discrimination in phonetic recognition. Zero-crossing rate Spectrum balance provides for the Characterization of fricatives an sibilants. High-resolution spectral information (i.e., pitch and format), LP parameters or filter-bank output provides formant and formant transition information. The patterns used in recognition may be represented either as time functions which span the entire word, or as a set of characteristic feature values which represent the subdivisions of the word. For example, in the system which characterizes the words by formant frequencies, each pattern may consist of a matrix of formant frequencies for a series of sampks over the duration of the word [7]. On the other hand, for example, in a digit-recognition system, each digit is segmented into initial. Medial, and final reasons based on the variation of Dower over time, and then the phonetic contents of each region are characterized using selected groups of features.

C) Feature Extraction :
Theoretically, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc [8].

The reason for computing the short-term spectrum is that the cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a nonlinear frequency scale (known as the Bark scale or the mel scale). This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. So, in the feature extraction, it is very common to perform a frequency warping of the frequency axis after the spectral computation.

D) Acoustic Models:
An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a phoneme. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes. An acoustic model is created by taking a large database of speech (called a speech corpus) and using special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models ("HMM"s). Each phoneme has its own HMM [9].

E) Language Models :
Approaches to language modeling can be divided in two major groups: deterministic (or grammar-based) and stochastic (or statistical).

Grammar-based language models are designed by experts on the basis of their knowledge of a language, function of a language model (LM) under development and intuition about the best way to represent linguistic entities and relations in a formal way. Within this framework a formal grammar of a language is constructed. The grammar is accompanied with a lexicon that specifies a list of possible terminal symbols. Probably the most popular type of grammars to use in language applications are context-free grammars (CFG) [10].

Statistical LMs emerge as a result of unsupervised LM estimation on a training corpus. A statistical LM usually starts as a set of void parameters that are estimated in the course of observation of language data. The choice of the parameters to estimate is still a matter of a designer's will. There are well known number of pro and contra as regards grammar-based and statistical approaches.

F) Modeling:
The objective of modeling technique is to generate speaker Models usng speaker specific feature vector. The speaker modeling technique divided into two classification speaker recognition and speaker identification. The speaker identification technique automatically identifies who is speaking on basis of individual information integrated in speech signal. The speaker recognisation is also divided into two parts that means speaker dependant and speaker

independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message .on the other hand in case of speaker recognition machine should extract speaker characteristics in the acoustic signal [5]. Speaker recognition can also be divide into two methods, text- dependent and text independent methods. In text dependent method the speaker say key words or sentences having the same text for both training and recognition trials.

3. Methods For Speech Recognition

TABLE 2: METHODS for SPEECH RECOGNITION

Methods	Specification
1. Hidden Markov Model	An alternative approach Speech recognition is to construct a statistical model of each word is the vocabulary and to recognize each input word as that word of the vocabulary whose model assigns the probability to the occurrence of the observed input pattern.
2. Deep Neural Network	A DNN is feed forward, artificial neural network that has more than one layer of hidden units between inputs and outputs
3. Dynamic Time Warping	DTW provides the time registration between each reference pattern and test pattern.

1. Hidden Markov Model:

Hidden Markov modeling is, as the name suggests, a modeling approach. Thus, there are three In addition the speech data can things to consider: the model, the method of computing the probability of the model giving rise to a particular output and the method of computing the parameters of the model from known examples of the word it is to represent. A hidden Markov model (HMM) is a doubly stochastic process for producing a sequence of observed symbols. An underlying stochastic finite state machine (FSM) drives a set of stochastic processes, which produce the symbols. When a state is entered after a state transition in the FSM, a symbol from that state's set of symbols is selected probabilistically for output. The term "hidden" is appropriate because the actual state of the FSM cannot be observed directly, only through the symbols emitted. In the case of isolated word recognition, each word in the vocabulary has a corresponding HMM. These HMMs might actually consist of HMMs that model subword units such as phonemes connected to form a single word model HMM. In the case of continuous word recognition, a single HMM corresponds to the domain grammar. This grammar model is constructed from word-model HMMs. The observable symbols correspond to (quantized) speech frame measurements [11].

2. Neural Network Model:

The neural network model that more commonly used and has the potentiating of speech recognition mainly include single layer perception model, multi-layer perception model, Kohonen self-organizing feature map model, radial basis function neural network, predictive neural network, etc. In addition, in order to make the neural network reflects the dynamic of the speech signal time-varying characteristics, delay neural network, recurrent neural network and so on [1].

A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs [12].

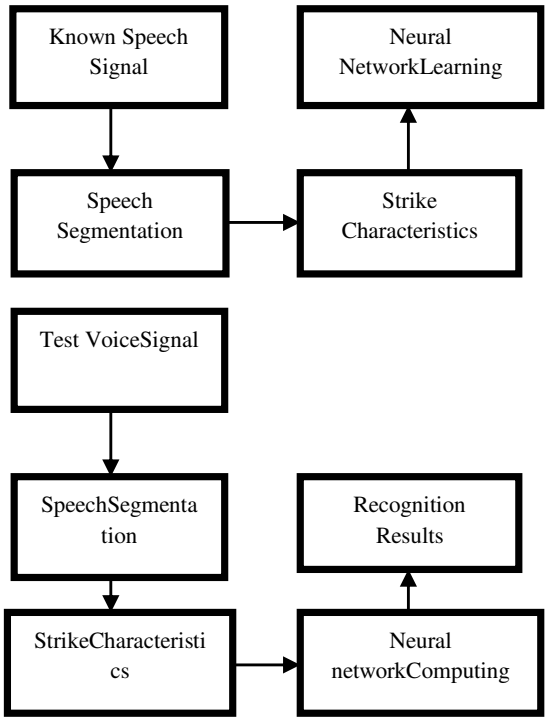

Figure 3: Neural network in speech recognition [1]

3. Dynamic Time Warping :

Although word recognition may be simplified to a linear-time-scaled word-template machine problem, such as those discussed above, the fact is that when a speaker pronounces the same word twice. The corresponding feature vectors or Patterns will never be exactly the same due to the variation in time, frequency and amplitude, thus, his necessary to align corresponding events in two different patterns, and the distance measure between the two patterns will be calculated according to this alignment. Dynamic Time Warping (DTW) provides the time registration between each reference pattern and the test pattern. This correspondence is determined by some user specified distance metric between a frame of the test pattern and a frame of the reference pattern. The optimum alignment function is determined such that It minimizes the total accumulated distance at each frame of the test pattern Thus, in this process, the time alignment and the distance computation are performed simultaneously After the scores for all the reference patterns are computed, the input is

identified as a word that belongs to the class of thereference token with minimum accumulated distance [2].

4. Application Of Speech Recognition

First, consider the resources. Text-based information retrieval is so useful and attractive because huge quantities of text documents are available over the Internet, and the quantity continues to increase exponentially due to the convenient access. For voice-based information retrieval, definitely multimedia and spoken content are the new trend, and such resources as rich as text-based resources can be realized even sooner given mature technologies. So this is not a problem at all [13].

Next, consider the retrieval accuracy. Clearly the accuracy for text-based information is acceptable to users and users even like it very much. In fact, the retrieval engines usually can properly rank and filter the retrieved documents which improve the perceived precision to a good extent. On the other hand, there are still serious problems with the accuracy of voice-based information retrieval, especially for spontaneous speech.

Speech under adverse environments in queries and/or target Documents which give very poor ASR accuracies. In fact, memory and computation requirements for voice-based information retrieval technologies also cause serious problems if a satisfactory accuracy has to be achieved. So the cost of memory and computation requirements is another problem coming together with the accuracy. Finally, consider the user-system interaction. For text-based information retrieval the retrieved documents are easily summarized on screen, thus easily scanned and selected by the user. The user can also select query terms suggested by the search engines for next iteration retrieval in an interactive process. Such convenient user-system interaction is actually a very important key which makes text-based information retrieval very attractive. For voice-based information retrieval, however, the situation is completely different. The multimedia/spoken documents are not easily summarized on-screen, thus difficult to scan and select [13].

TABLE 3: COMPARITION BETWEEN VOICE-BASED AND TEXT-BASED INFORMATION RETRIEVAL [13].

	Text based	Voice based
Resources	Rich resources huge quantities of text documents available over the internet. Quantity continues to increase exponentially due to convenient access.	Spoken/Multimedia contents are the new trend. Can be realized even sooner given mature technologies.
Accuracy	Retrieval Accuracy acceptable to users. Retrieving documents properly ranked and filtered.	Problems with speech recognition errors, especially for spontaneous speech under adverse environment.
User-System Interaction	Retrieved documents easily summarized on screens, thus easily scanned and selected by users. Users may easily select query terms suggested for next iteration retrieval in an interactive process.	Spoken/Multimedia documents not easily summarized on screen, thus difficult to scan and select. Lacks efficient user-system interaction.

A. User interaction for telecommunications services: Traditionally, user interaction, such as designation of the called party for telephone calls was provided through operator assistance. Today, most of it is automated by the use of the DTMF input. Replacement of DTMF by voice recognition improves the human interfaces, making it user-friendly as the traditional operator assistance, and sometimes significantly enhances the service capability [14].

B. Automation of information provision Directory assistance to assist finding the telephone number of the target party is a typical example of voice recognition application. There are also many applications of voice recognition already realized in the field of information retrieval, e.g. stock price in the exchange market and balance in a bank account, through a telephone network [14].

C. Departments in major U.S. telecom operator Sprint PCS have the largest digital wireless network at the same time known for excellence and innovative customer service. The opening voice-driven systems for clients since 2000. The system provides customer service, voice dialing, check number, and change addresses and other services. In addition, China Telecom has launched a voice recognition integration of value-added services system CELL-VVAS, (VOICEVALUE-ADDED SYSTEM), the system uses a distributed excellent recognition engine, developed a stable and efficient application. The system also perfectly integrated telecommunications switching network application to provide users with a variety of user-friendly, personalized service [1].

5. Comparative Analysis Of Speech Recognition

Some benchmark works done in the past are presented in Table 4. However, the results they reported, although quite encouraging on most occasions, were obtained using only a selected number of speech classes in their experiments. This leaves a question that how these speech features will perform when applied to speechs other than those considered in their work.

Authors	Feature of Speech	Classifier	Task/Search Type	Systems	Performance
1. Joseph Picone [15]	Continuous speech	HMM	Speaker dependent	Tangora System	95-99%
			Speaker Independent	SPHINX System	84-96%
			Multiple DSP	HPCDR System	42-92% (No. Of Mixture)
			Single Multivariate Gaussian Distribution	Improved Digit Recognition	30-44% (No. Of Models)
2. Geoffrey, Li Deng [12]	Large Vocabulary Continuous Speech	HMM	DNN	Google Voice Input	88%
				YouTube	53%
			GMM	Google Voice Input	84%
				YouTube	48%
3. Todd A., Herve [16]	Continuous	HMM/ANN	DBN	Clean System	90-91%
				Noisy System	70-90%
			HMM/ANN	Clean System	70-80%
				Noisy System	30-75%
4. Nihat , Ulvi [17]	Compared with Voice Command Input	Microprocessor	Zero Signal for Count & Time	PIC18F452 Microcontroller	65-70%
5. E. Chandra, C. Sunitha [4]	Continuous Speech	RBF NNs	Classical method of Classification	Minimal Eigen Values	97.73%
			Burg's Model	Identify Right Speaker	94.82%
6. Jose, David [18]	Impaired People Speech	Fourier Transform	MATLAB	Sagebrush System	10 Recordings per Word
7. Olli, David [19]	Noise Robust Speech	Single Mixture HMMs	Normalized Feature Vector	Speaker Dependent System	83.7%
8. Yifan Gong [20]	Speech in Noisy Environments	HMM Vector Equalization	10dB Gaussian Additive Noise	Multi Speaker System	98.3%
9. Steve, Herve [21]	Connectionist-statistical speech	CI-HMM	Connectionist Probability Estimators	Context-independent MLP-HMM hybrid system	92%
10. Ossama, Abdel [22]	Voice search large Vocabulary speech	CNN	84 Feature Maps per Section	Two Hidden Layers System	92%
11. Alin, Jacek [23]	Visual and auditory Information for speech	PCA	Relevant Motion Information	Static Visual Feature	81.11%

A brief breakup of the speech recognition works which done in the past by a number of researchers is given in table 4 as:

Joseph Picone [15] presented a continuous speech recognition system using Hidden Markov models, and the authors have formed four systems for the experimental purpose is given as:

1. First, The Tangora System, a speaker dependent isolated utterance, speech recognition system scalable from 5,000 words to 20,000 words was the product of IBMs long-term commitment to applying stochastic modeling to speech recognition. Performance has been measured across a variety of large vocabulary recognition tasks from 95 to 99 %.
2. Second, The SPHINX System is a speaker independent continuous speech recognizer based on triphone acoustic models (sequences of three phones). Application of SPHINX to the DARPA Resource Management task (a language that consists of a 1000 word vocabulary and a finite state automaton of over 7,000 nodes and 65,000 arcs). Word accuracy of the SPHINX system as a function of the recognition unit of the DARPA Task Domain database from 84 to 96%.
3. Third, HPCDR (High Performance Connected Digit Recognition), One derivative of this work was a demonstration of a real-time HMM digit recognition system using multiple Digital Signal Processors (DSPs) on the ASPEN multiprocessing system. Sentence accuracy as a function of the number of mixtures per state in the case of 10 states per model with one model per digit. Here Performance of System from 42 to 92 %.
4. Fourth, IDR (Improved Digit Recognition). Her System used a single multivariate Gaussian distribution per frame in the reference model to improve the performance of speech recognition. Sentence accuracy from 33 to 44 % based on number of Models.

Geoffrey, Li Deng [12] presented Large Vocabulary Continuous Speech feature using Hidden Markov models. Here speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. It shows that DNN-HMMs consistently outperform GMM-HMMs that are trained on the same amount of data, sometimes by a large margin. For some tasks, DNN-HMMs also outperform GMM-HMMs that are trained on much more data. Here used two systems with HMM.

1. First, for Google Voice Input and YouTube for Both DNN-HMM and GMM-HMM Separately. For Google Voice Input in DNN-HMM, Performance accuracy is up to 88% and for YouTube in DNN-HMM, Performance accuracy is up to 53%.
2. Second, for Google Voice Input in GMM-HMM, Performance Accuracy is up to 84% and YouTube in GMM-HMM, Performance accuracy is up to 48%.

Todd A., Herve [16] presented Auxiliary Continuous Information Feature using HMM/ANNs model.

Here investigate different approaches to incorporating this auxiliary information using dynamic Bayesian networks (DBNs) or hybrid HMM/ANNs (HMMs with artificial neural networks). Auxiliary features that have a dependency on the state can be referred to as mid-level auxiliary information (e.g., articulatory features) while auxiliary features that do not have a direct dependency upon the state can be referred to as high-level auxiliary information. Since it is not always clear with some features whether they contain mid-level or high-level information. In DBNs, Results using the three types of auxiliary features of ROS (Rate of Speech), pitch, and energy.

Both DBNs and HMM/ANNs use two types of speech system Clean and Noisy to conclude the Performance of the system.

1. First, for DBNs, In Clean system accuracy is from 90 to 91% and In Noisy System accuracy from 70 to 90%.
2. Second, for HMM/ANNs, In Clean System Accuracy from 70 to 80% and In Noisy System accuracy from 30 to 75%.

Nihat, Ulvi[17] compared with Voice Command using a PIC18F452 microcontroller as Microprocessor Based Voice Recognition system model.

In this, using a PIC18F452 microcontroller, two registered words, compared with a voice command input. Recognition of the audio signals zero crossing for the count and time was examined. All the speech sounds are composed of linear combination of sine waves with different frequencies.

The frequency values of human voices are ranged between 300Hz-3300Hz. According to the Nyquist Theorem, it is possible to make an effective sampling with a double frequency value of sound frequency and a higher sampling frequency. Here, the voice obtained from an electrets microphone is upgraded by LM386. In this study, PIC18F452 is used as a microprocessor. In this, PIC18F452 is used as a microprocessor. PIC18F452 powerful 10 MIPS (100nanosecond instruction execution) yet easy-to-program (only 77 single word instructions) CMOS FLASH-based 8-bit microcontroller packs. The PIC18F452 features 256 bytes of EEPROM, Self programming, 2 capture/ compare/PWM (10 bit) functions, 8 channels of 10-bit Analog-to-Digital (A/D) converter, the synchronous serial port can be configured as either 3-wire Serial Peripheral Interface or the 2-wire Inter-Integrated Circuit bus and Addressable Universal Asynchronous Receiver Transmitter.

It is possible to perform 65-70% of voice recognition in trials.

E. Chandra, C. Sunitha [4] presented Speech and Speaker identification using Neural Networks.

Here for classification, probabilistic and radial basis functions (RBF) NNs have been used.

Two simple methods for classification chosen are classical and neural-based ones. Both methods have their input data from the TM minimal Eigen value algorithm or from the Burg's curve directly.

If we have 20 speakers with 11 digits means we have 220 different classes. At the stage of classification, we have 11 classes for voice recognition (digits from zero to ten) and 20 for speaker identification (the number of speakers in the author's base).

Here two experiments have performed by authors as:

1. First, Classification with the application of the TM Minimal Eigen values Algorithm, however, has increased its performance rate to 97.73%.
2. Second, Burg's model with Conventional speech classification. In this the rate of identifying the right speaker successfully, has reached 94.82%.

Jose, David [18] performed recognition of Impaired people's voice using Fourier transform.

In this, the authors have shown how MATLAB can be used to realize for Speech Recognition. MATLAB has the flexibility to implement complex algorithms for digital processing of signals. Here investigation is based in a strong way, in the Fourier Transform. To represent the signal in the frequency domain has used the DFT. A way to apply the DFT, is using the FFT. The record of the phonemes was realized with the program Recall Version 2.4a de Sagebrush Systems. Each one of the phonemes was recorded 10 times. Of the recordings they chose the best five to analyze it. These five recordings are those ones that presented less noise. To the 5 recordings of each phoneme they applied a program made in MATLAB, with this program, we obtained the spectrum of the filtered signal in different bands. Finally, from these spectra they picked the most common and they call it characteristic standard of the phoneme.

Finally, the authors took new voice recordings to probe the system, they used 10 recordings of each word; the results are next As a final adjustment.

Olli, David and Kari [19] presented Noise Robust speech recognition using Single mixture HMMs.

In this, the authors were evaluated the recursive feature vector normalization approach in speaker-dependent name recognition and speaker independent connected digit recognition tasks. In the tests, 13 MFCCs (including the zcrorh cepstral coefficiect), their first and second order time dematives were extract& from the incoming signal. The results obtained in speaker-dependent name recognition with various values of ***N*** at different SNRs. Clearly, one has to buffer at least 20 feature vectors (0.2 secs.) in order to achieve a good recognition performance at low **SNRs** as well.

The greatest absolute performance improvement was achieved by authors with single mixture HMMs in the most noisy conditions is 83.7%.

Yifan Gong [20] presented Speech recognition in Noisy Environments using HMM vector equalization.

In this the essential points in noisy speech recognition consist of incorporating time and frequency correlations, giving more importance to high SNR portions of speech in decision making, exploiting task-specific a priori knowledge both of speech and of noise, using class-dependent processing, and including auditory models in speech processing. In a noisy environment, the distribution of parameters that give very good recognition results for clean speech can be very sensitive to disturbances, which introduce a mismatch between training and testing conditions.

Here, authors result 10 digits vocabulary size using Hmm vector equalization with multi speaker gives a performance with accuracy 77.3% in noisy environments compare to clean environments.

Steve, Herve [21] presented Connectionist-statistical speech recognition using Hidden Markov models.

Here, the performance of such a system using a multilayer perceptron probability estimator evaluated on the speaker-independent DARPA Resource Management database. In this, they are concerned with building statisticalmodels of the speech signal in the feature vector domain. They use a set of basic HMM's, corresponding to phones. These are concatenated or built into networks, to form words and sentences, according to the lexicon and language model.

The context-independent MLP-HMM hybrid system had a word accuracy is 92%. They have been found the connectionist HMM framework a good one in which to explore issues such as robust front ends, speaker adaptation, consistency modeling, and context dependent phone modeling.

Ossama, Abdel [22] presented Voice search large Vocabulary Speech using Convolutional Neural Network models.

Here, the authors have been shown a further error rate reduction can be obtained by using Convolutional neural networks (CNNs) and proposed a limited-weight-sharing scheme that can better model speech features. In this, Experimental results show that CNNs reduce the error rate

by 6%-10% compared with DNNs on the TIMIT phone recognition and the voice search large vocabulary speech recognition tasks. The experiments of this section have been conducted on two speech recognition tasks to evaluate the effectiveness of CNNs in ASR: small-scale phone recognition in TIMIT and large vocabulary voice search (VS) task. The DNN had three hidden layers while the CNN had one pair of convolution and pooling plies in addition to two hidden fully connected layers. The CNN layer used limited weight sharing and had 84 feature maps per section. The CNN improves performance by about an 8% relative error reduction over the DNN. Thus, the performance achieved by up to 92%.

Alin, Jacek [23] presented Visual and auditory Information for speech recognition using Principle Component Analysis model.

The current audio-only speech recognition still lacks the expected robustness when the Signal to Noise Ratio (SNR) decreases. The video information is not affected by noise which makes it an ideal candidate for data fusion for speech recognition benefit. Here the authors have shown that most of the techniques used for extraction of static visual features result in equivalent features or at least the most informative features exhibit this property. They show that the audio-video recognition based on the true motion features, namely obtained by performing optical flow analysis, outperforms the other settings in low SNR conditions. The most popular method for this is Principal Component Analysis (PCA).

Other methods which were used as an alternative to PCA are based on the discrete cosine transform and discrete wavelet transforms. Here, the accuracy of Word recognition percentage for static visual features in clear audio with equal weights is up to 81.11%.

6. Conclusion

In this paper, given a review of Speech recognition. The area of Speech recognition is continually changing and improving. Speech recognition technology is capable to make possible to communicate with disabled persons. It makes control of digital system. In future, vast possibilities to enhance the area of speech recognition technology. By enhancing of speech recognition can provide better services for disable persons. Speech recognition can provide a secure environment to our system by voice authentication. Different methods and their accuracy also tabulated that shows the use of HMM and ANN model is much wider used methods for continuous speech recognition process. In the future, the correctness of speech recognition and the quality of speech will be more improve that's makes communication so easy and reliable for everybody including disable persons. Future systems must be more efficient and capable compare to traditional systems.

Future scope: The world of Speech recognition is rapidly changing and evolving. Early applications of technology have achieved varying degrees of success. The promise for the future is significantly higher performance for almost every speech recognition technology area, with more robustness to speakers, background noise etc. This will ultimately lead to reliable, robust voice interfaces to every telecommunications service that is offered, thereby making them universally available.

REFERENCES

[1] Jianliang Meng, Junwei Zhang and Haoquan Zhao, "Overview of the Speech Recognition Technology", 2012 Fourth International Conference on Computational and Information Sciences, 978-0-7695-4789-3/12$26.00©2012 IEEE.

[2] Andress S. Spanias, Frank H. Wu, "Speech Coding and Speech Recognition Technologies: A Review", CH3006-4/91/0000-0572$1.000 IEEE.

[3] Jeff Zadeh, "Technology of speech for a computer system", DECEMBER 2003/JANUARY 2004, 0278-6648/03/$17.00 © 2003 IEEE.

[4] E. Chandra and C. "A review on Speech and Speaker Authentication System using Voice Signal feature selection and Extraction", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[5] Santosh K.Gaikwad, Bharti W.Gwali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[6] Lawrence R. Rabiner, "Applications of speech recognition in the area of telecommunication", 0-7803-3698-4/97/$10.00 *0* 1997 IEEE.

[7] Tingyao Wu, D. Van Compernolle, H. Van hamme, "Feature Selection in Speech and Speaker Recognition" June 2009. U.D.C. 681.3_I27. Phd Thesis.

[8] Urmila Shrawankar, Vilas Thakar, "Techniques for Feature Extraction in Speech Recognition System : A Comparative Study".

[9] Chris Biemann, Dirk Schnelle-Walka, "Unsupervised acquisition of acoustic models for speech-to-text alignment", Master-Thesis von Benjamin Milde 10. April 2014.

[10] Maxim Khalilov, J. Adri´an Rodr´ıguez Fonollosa, "New Statistical And Syntactic Models For Machine Translation", TALP Research Center, Speech Processing Group, Barcelona, October 2009.

[11] Richard D. Peacocke, Daryl H. Graf, "An Introduction to Speech and Speaker Recognition", Bell-Northern Research,IEEE August 1990.

[12] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition", Digital Object Identifier 10.1109/ MSP.2012.2205597,Date of publication: 15 October 2012.

[13] Lin-shan Lee and Yi-cheng Pan, "Voice-based Information Retrieval How far are we from the text-based information retrieval?", IEEE ASRU 2009.

[14] Masanobu Fujioka, Seiicbi Yamamoto, Naomi lnoue, Makoto Nakamura and Takashi Mukasa, "Experience and Evolution of Voice recognition applications for telecommunicati0ns services" 0-7803-4984-9/98/$10.00 *0* 1998 IEEE.

[15] Joseph Picone, "Continuous Speech Recognition Using Hidden Markov Models", IEEE ASSP MAGAZINE JULY 1990.

[16] Todd A. Stephenson, Mathew Magimai Doss and Hervé Bourlard,"Speech Recognition with Auxiliary Information", IEEE transactions on speech and audio processing, vol. 12, no. 3, May 2004.

[17] Nihat Öztürk and Ulvi Ünözkan, "Microprocessor Based Voice Recognition System Realization", 978-1-4244-6904-8/10/$26.00 ©2010 IEEE.

[18] José Leonardo Plaza-Aguilar, David Báez-López, Luis Guerrero-Ojeda and Jorge Rodríguez Asomoza, "A Voice Recognition System for Speech Impaired People", Proceedings of the 14th International Conference on Electronics, Communications and Computers (CONIELECOMP'04) 0-7695-2074-X/04 $ 20.00 © 2004 IEEE.

[19] Olli Viikki**,** David Bye and Kari Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", 0-7803-4428-6/98 $70.08 0 1998 IEEE.

[20] Yifan Gong, "Speech recognition in noisy environments: A survey", Speech Communication 16 (199.5) 261-291,0167-6393/95/$09.50 0 1995 Elsevier Science B.V.

[21] Steve Renals, Nelson Morgan, Herve Bourlard and Michael Cohen, "Connectionist Probability Estimators in HMM Speech Recognition", IEEE Transactions on Speech and Audio Processing, VOL. 2, NO. 1, PART 11, JANUARY 1994.

[22] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 22, NO. 10, OCTOBER 2014.

[23] Alin G. Chit¸u, Leon J.M. Rothkrantz, Pascal Wiggers and Jacek C. Wojdel, "Comparison between different feature extraction techniques for audio-visual speech recognition", Journal on Multimodal User Interfaces, Vol. 1, No. 1, March 2007.