CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 EVOLUTION OF SPEECH RECOGNITION

Amongst the cognitive senses, Speech is the most effective channel. Speech is most initial communication medium of transferring information, emotions & thoughts. Moreover, it is also a unique medium of passing human intellect from one member to another [31]. In science, the human voice has long been an accepted feature. Bill Gates (co-founder of Microsoft Corp.) in his book 'The Road Ahead' describes that speech recognition [65] process as a very important tool or innovation for future computing. In human computer interaction, the speech [58] identification is always looked like an allure. Speech signal identification is a way of change in an acoustic speech signals taken through a microphone or a telephone, in set of words. The identified words can be the last output for different applications like commands of signal and control, raw material or signal, document and penetrate work preparation. The identification of speech signal has been a major area goal for researchers for more than four decades. Like a machine, by allocators in whole environment, the spoken digits are very specific on very human being [89] [85]. Classification and recognition of signals together involves a man's perception. The excellent example of human disposition may be arrival of language to classify inherent and identical sequence. Discovering and recognizing patterns in the speech signal is a toughest task for the sequence recognition machine. Speech [1] [2] recognition process includes complex structure which consists of not only the passing of voice but also the gestures, the language, the subject and the capability of destination. The working of a speech recognizer system depends on its task and its architecture. People have the capability to communicate with machine and the

computer through keyboards or other input devices which are slower in working. The most important and the easiest way for this communication is speech. It is clear that computer is not able to understand what signal has been uttered. But, it is controlled via speech signal. The aim of research on Automatic Speech Recognition (ASR) [73] is to build machines i.e. different from human's ability to communicate with in natural spoken language. In this way, speech recognition is aptly a major science.

A speech recognition system requires front-end signal processing which changes the speech signal to represent values and it is shown for next analysis and working. Parametric representation of speaking signal may include zero crossing rates, in short time capability of doing work (energy), level crossing rates, and other related representation of values. By Feature extraction technique we can convert a pattern of spoken samples at the front end into a set of observation vectors that show a probabilistic space in events for the category of signal where it works. Feature extraction portion is complex for speech identification. Features are capturing dynamic & robust spectral fact in a given spoken signal. They are very hard to find out from speech signal. With the speech recognition process, feature extraction problem is a difficult task. There is a great need to explore new ways for the said problem. In real world observable results are known as signals. The speech indication may be discontinued in behaviour (e.g., character). The basic theory of Hidden Markov Model (HMM) [2] was published in a series of papers. However, widespread knowledge and application Theory of HMMs to speech signal working has occurred only last few years. There may be many reasons behind this i.e. the case of communication. First, the basic matter of secret Markov presentation was introduced in mathematical papers. Basically it was read by researchers working on difficulty in speaking signal working also, there after the technique could not give sufficient learning data for researchers to get the basics and to be capable to apply an approach to their research. Now a days we have access to many tutorial papers which gives us sufficient level of detail to start research work using HMMs for individual speech signal processing applications.

1.2 AUTOMATIC SPEECH RECOGNITION

Speaker identification systems are commercially available with limited capabilities. In the last few decades, the working of these systems have taken grand leaps, as a different function of the one form to another form of energy conversion (microphone to telephone), orator (orator independent to orator dependent) and operating environments. The basic task of (ASR) is to figure out a pattern of words from a stream of acoustic data.

To understand Automatic Speech Recognition that is producing actions for speaking: ASR systems includes some large component as shown in diagram1.1, the very first block includes the acoustic environment plus the device such as microphone, current modulator etc A/D converter etc. The second block is the feature extraction sub system, also named as the front end signal processing. It derives acoustic representations. The succeeding two blocks in diagram1.1 demonstrate the core acoustic arrangement that is equivalent [5] procedure of speech identification. Virtually all ASR systems are the illustration of speaking [6], like Cepstral illustration is worked out over successive intervals. These demonstrations of speaking structures are then associated with the spectra for speaking that were used for drill. These all contrast can be observed as a native match. The inclusive match is examined for the better signal pattern of words, and is firm by integrating many native matches. The limited context cannot create a single rigid choice of the adjoining speaking class, but it produces clusters of distances equivalent to possible sounds: A worldwide search to get a closest pattern of speaking classes. In Figure 1.1, fifth block language model, provides hypotheses that is used in the global search for the speech. Language model moreover can give the worldwide translator result for remaining occurring processing.

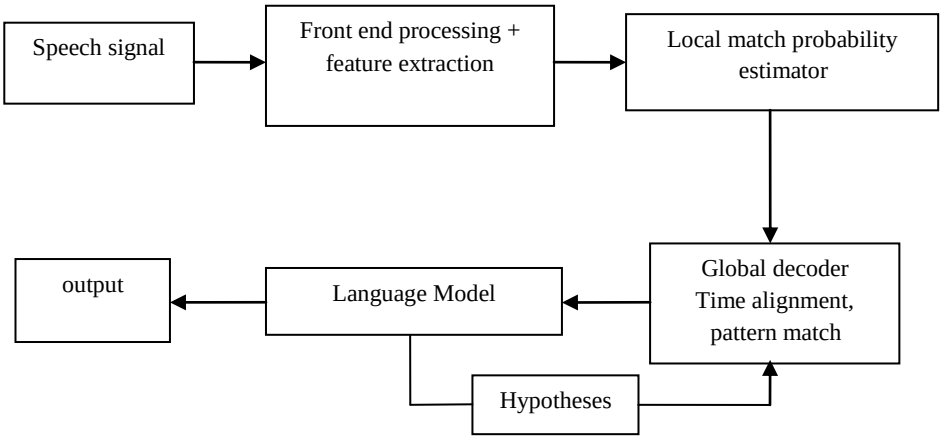

Figure 1.1: Block diagram for Speech Recognition

If the decoder produces the N greatest prospective sentences but not the most prospective sentence this block can get these sentences according to grammar or semantic.

The human speech fabrication and insight apparatus provides insight in the training material for the altered feature extraction techniques. Communication is made by the pulsation in vocal cords when breath is respired from the lungs. We can get the assortment of communication sounds because of the disparity in mass and the length of vocal cords and the vocal tract of auriculars. The length of the men's vocal cords is 17 - 24 mm and an average essential frequency of 125 Hertz, but the length of the women's vocal cords is 13 - 17 mm and an average essential frequency of 200 Hz. It is not clear that how the brain understands communication signals. The human ears sensitivity to speech is non-linear. A non-uniform Mel scale explained as pitch of a 1000 Hz tone in research [89].

1.3 MOTIVATION OF RESEARCH

Speech identification is always observed as an important field in human computer interface [85]. A speech identification system consists of algorithms from the variety of disciplines, like statistical sequence identification, matter of communication, signal sensation, linguistic, etc. The man-machine interface is the basic or primary focus for research in speech processing. One of the top most important aims of carrying out exploration in speech signal identification through machine is inter disciplinary behaviour and the capacity to solve the problem of scientists.

A major and motivational part for selecting Hindi as the language for our gratitude system i.e. comes from its native application because this is our mother tongue. Out of India, people in several other countries like Mauritius, Singapore, Nepal, Fiji, Guyana, Suriname, Trinidad, UAE, etc. can easily understand and even speak Hindi.

Actual-time communication recognizers can have extensive uses in numerous presentations, containing simplified man-appliance signal, support for hearing impaired entities and the physically spiked peoples, telephone additional worker, and other man-apparatus edge errands. The insulated word identification systems [65] [32] were amongst the first speech identification systems built due to relatively forthright method. The theory and implementation of Automatic Speech Recognition [65] [21], in the form of Isolated Digit Recognition (IDR) system is discussed in this research. The aim of automatic communication identification is to generate methods and systems for initial computers systems that accept communication signal input, Speech signal recognition evolve like as speech signal-to-text conversion problem, speaker recognition, isolated word recognition etc... Users wanted their speech signals, to be recognised by system. We have made important development in (ASR) intended for efficiently-well defined presentations such as notation and average terminology operation working chores in both environments by use of some elementary features for example, Linear Predictive (LPC) [22], Perceptual Linear Prediction (PLP) [40][29] and Wavelets [1]. Fiber Optic

communication is a cradle for creating major developments in great noisy surroundings. In this research, our goal is to get the advance techniques.

1.4 DESIGN PROBLEM

ASR systems has most successful basis like HMM. ASR performance with HMM is ideal for mobile distance from commencing human-like identification for composite tasks, such as, a huge terminology or regular impulsive situations. There is a lot of undesirable variability in consistency in speaking signal. These dissimilar sources of inconsistency are:

Humans introduce different variability while speaking due to different reasons. Well-communicated speaker might adjust their reactions. The orator changeability can be auxiliary classified as with-in recite, anywhere; the previous foundation of inconsistency defends intra-orator changeability and the future as inter-speaker unevenness.

Medium and transducer: Transducer is a device which converts one form of energy into another form. The speech signal or source signal can be composed by various transducers and can be communicated by dissimilar networks, for example microphone speech and telephone speech. For illustration, the telephone speech band is restricted among 300 Hertz and 3300 Hertz, where the microphone speech has a larger bandwidth. So, throughout the speech process in the broadcast networks the power speech signals have the lot of changeability.

***Atmosphere*:** The involvement of the transducer, identifying the speech signal, it is not only the aural density wave segmented by the speech manufactured system of envisioned speaker, also the signal comes from the contiguous environments: i.e. noise along with different analog signals. The uproar signal inhibits by the speech signal ensuing in supplementary patchiness.

Phonetic of speech: The audio recognition of the phonemes is extremely reliant upon the contiguous phonemes. For illustration, the acoustic realization of phoneme /a/ in word cat /k/ /a/ /t/ and word bat /b/ /a/ /t/ is dissimilar. This is generally owing to co-articulation. In other words, we can say that Co-articulation is "Covering of contiguous articulations". Variability contemporary in the speech indicator is stimuli in the ASR structure at numerous stages. For example, the broadcast frequency

variability has a consequence on the aural chin vectors, whereas, the articulation disparity has upshot upon both the aural feature vector and the verbal prototypical. Supplementary foundations of information also deliver extra material to condense inconsistency.

1.5 APPLICATIONS

The range of Speech Recognition applications is wide and includes:

- Military Services
- Spoken Audio Search
- Health Care Department
- Voice recognition Aid for Disabled People
- Automatic Speech Recognition systems
- Direct Voice Input in airplane cockpits
- Home Automation
- Voice Controlled Appliances
- Wireless communication
- In Robotics
- Medical field

1.6 AIMS OF THE THESIS

The foremost ideas of the thesis:

(a) Develop robust speech recognition system for Hindi language that should work in all the environments.

(b) To develop hybrid features for Hindi speech recognition which should give better recognition efficiency than the existing features.

(c) To prepare proper Hindi speech database to be used in audio and visual speech recognition experiments.

(d) To see the recognition performance of existing LPC, PLP besides wavelet grounded features for Hindi speech system.

(e) Use of (HTK) Hidden Markov Model Tool kit software for evaluating the performance of hybrid features.

(f) Development of audio-visual features for robust Hindi speech recognition in all environments.

1.7 THESIS OUTCOMES AND CONTRIBUTIONS

The following are contributions and outcomes of the thesis:

(a) **Database Preparation:** Proper database is required for Hindi speech recognition. Due to non-availability of a suitable Hindi database, we have prepared three different sets of Hindi speech databases for evaluating the presentation of Hindi communication recognition organizations in fresh and noisy environment.

(b) **Implementation of feature extraction methods:** In this thesis some basic feature extraction methods used for other languages have been studied and implemented to study the behavior of these methods in Hindi speech recognition environment.

(c) **For Hindi speech Proposed method for evaluating modified hybrid features in HTK software engine:** In this work three modified hybrid features studied are Perceptual Linear Coefficients, Bark frequency Cepstral coefficients and Mel frequency-perceptual linear prediction [19][59]are evaluated using HTK Tool. For doing so, new lexicon and grammar is prepared and simultaneously new word (.mif) files are created. By changing the environment for Hindi speech recognition, it is found that HTK software engine is giving almost 99.9% accuracy in case of clean database and also performs better in case of noisy conditions as compared to other simulation results.

(d) **For audio visual Hindi speech recognition, Proposed various methods for extracting visual features:** Different existing methods for lip localization introducing necessary desired modifications are made and tried. A new

method for localization of lips is proposed as a hybrid of two methods change in colour system method and colour intensity. Visual features are extracted from a database, uttering stops like B, P, K and G. In this work we are capable to demonstrate that in noisy environment recognition exactness is increased in case of phoneme-viseme recognition as compared to only phoneme recognition.

1.8 THESIS ORGANIZATION

Brief explanation of following chapters:

First Chapter: This chapter is dealing with the brief literature review of speech recognition. Brief Review of speech recognition for Indian languages is also mentioned in this chapter.

Second Chapter: In this chapter the classification techniques like HMM, HTK and Linear Discriminant Analysis (LDA) used for Hindi speech recognition. There are different classifiers available and those can be used for the purpose of speech recognition. In ASR investigation an accurate merit has been found among recognition, classification. In the given thesis, subdivision of insulated Hindi digits has been previously done at the time of file preparation. Hence, in all the works presented in this thesis, the ASR task is limited to isolate Hindi digits classification.

Third Chapter: In this chapter discussion is about various acoustic feature extraction techniques used in this work. In feature extraction, spoken signal is changed into the set of feature vectors that contains the signature features about certain utterance which is important for recognition. The features extraction represents the most significant property in the spoken signal. This chapter starts with explaining basic feature extraction methods like Mel Frequency Cepstral Coffecient (MFCC), PLP, & LPC. Research has also explained wavelet based features which are advanced feature. At the last, the process of extracting three newly modified hybrid features like Bark

Frequency Cepstral Coffecient (BFCC), MF-PLP coefficients and Revised Perceptual Linear Prediction (RPLP) features are explained.

Fourth Chapter: This chapter defines the experimental setup and evaluation results for acoustic features. This chapter starts with explaining the method of preparation of different databases used in this research. The databases were prepared at Rajkumar Goel Institute of Technology Ghaziabad, and at Aligarh Muslim University, Aligarh, India. For preparation of databases, speakers from different region of India were prepared mostly in the age group of 18-27 ages. Artificial noisy database remain prepared for Hindi numbers via tallying altered categories of noises from NOISEX-92 database to fresh Hindi numbers database. Then it deals with the experimental setup and results of different experiments conducted on these databases for extracting different features. We have executed the system by MATLAB and HMT in Linux environment. In the initial phase of the research, feature extraction and classification using LDA and HMM is done using MATLAB programs. MATLAB and HTK software is used for the evaluation of Different features. These toolkit goals at construction and operating HMM. MATLAB and Hidden Markov Toolkit also used for the evaluation of the performance of three newly modified features (BFCC, RPLP, and MF-PLP) with basic experiments.

Fifth Chapter: This chapter is dedicated to audio visual feature extraction techniques and their results. The procedure of graphic features in AVSR is inspired through speaking signal construction device & ordinary capability of humans to moderate acoustic uncertainty by visual signals. The visual data gives us balancing features that cannot be degraded via the audio noise of the atmosphere. Our main focus of the work is on the recognition of ends; because of audio-based recognizer with less correctness. A new method for localization of lips is proposed as a hybrid of two methods i.e. Color Conversion method and Lip Color intensity method. In determination of exact lip boundary from the binarized image, new methods have been tried and tested. Under the provided conditions of the video recordings, comparison of existing and proposed methods has been done.

Sixth Chapter: This chapter completes the thesis through summary of the work. For Hindi speech recognition the modified hybrid features are suitable. These robust

features make it suitable for voice applications, military services, voice recognition aid for disabled persons etc. Presently, there is very little research in the area of Hindi speaking recognition. In this research it is found that newly modified hybrid features (BFCC, RPLP, and MF-PLP) provide better recognition accuracy as compared to existing basic features specially in case of noisy conditions. As expected, the proposed methods for extracting visual features are performing better than the existing methods.

1.9 GENERAL REVIEW OF LITERATURE

Computer,-man interaction is explained in previous concerned literature about methods that interrelate with computers. Maximum users can interrelate with the computer by many traditional techniques of console and mouse. The CCTV of computer is output device. Speaking recognition schemes helps the peoples by one method or other who is not able to use many conventional peripherals. In the simplest form, machine includes two subsystems, such as spontaneous speaking recognition & verification. The objective of ASR is to record normal speech to know the sense of the transcript. So speech communication process might be taken into mind as all adjustable information.

System performance for speech applications is controlled via different ASR methods. Project Engineer's work provides the thoughtful recognition in the speech field. This is applicable to get actual precise domain for the prompt communication.

***Different Types of ASR*:** The first speech recognition system appeared in 1952. It consists of a device that can recognize single spoken isolated digits. Another early device was IBM shoebox. In the 1960, laboratories of some other countries innovated the efficiency of building singular purpose hardware perform i.e. speaking recognition job: but initially systems were luxurious and the hardware devices could identify only some isolated numbers. In the late 1990s, the software was developed for desktop transcript i.e. accessible for a few dollars. Mainly the goal of Direct Voice Input (DVI) devices is voice control & command, but the voice-based document creation performed by Large Vocabulary [48] Continuous Speech Recognition (LVCSR) systems

was also introduced. Both are similar technology. DVI schemes generally respond immediately to voice appreciation.

Detailed instances of presentation of ASR may contain but it is not limited to the following Telecom assistants - aimed at personal management systems plus repertory dialing

(a) Factory management Process Management - for quality control, stocktaking, and measurement.

(b) Response of cooperative voice- Data services like standard market estimations.

(c) Large vocabulary –For medical services.

***Speech Recognition Techniques*:** These can be categorized in the following manner: -

(a) Methods based on information: Human information about dissimilarities in speaking signal is not machine implicit into method.

(b) Template based equivalent method [33]: New speech is linked beside a set of pre -recorded template in order to find out excellent match. In the research change in real time is a difficult method [34]. In this method, corresponding words having the patterns that contain representative word feature vectors pattern. For a word the lowest scoring path recognizes the optical alignment sequence and term pattern finds out by the lower most score of standard sequence of words.

(c) Statistical method: Here speech samples are modeled statistically. For this purpose, research is using automatic and statistical method, basically the HMM. In the modern ages of research, a conversational speech recognition has come up as new approach to the challenging problems, allotment of aptitude to overcome some traditional HMM limitations [34] [66]. Tremendous dissimilarity in the visible audio information of mostly co-articulated impulsive speaking signal, huge number of formless Gaussian mixture mechanisms [73] have been developed which delivers a ironic building for the partly observation.

(d) Artificial intelligence approach: Recognition procedure attempts to mechanize according to the human knowledge in imagining, examining, and lastly making a result in the evaluated audio features.

(e) Knowledge based methods: These methods come out from demerits of traditional HMM; machine learning approaches could be introduced like genetic algorithm programming and neural networks.

1.10 AUTOMATIC SPEECH RECOGNITION (ASR) REVIEW

Here the current trends in speech were reviewed. Communication was introduced via the speech signal was set in a highly composite method [14]. It took till the early 1970s when Lenny Baum innovated new HMM method to speaking recognition. The HMM method was the most useful application. By late 1970's, the research inclined from towards remote word recognition.

For this work a building level is used & the fresh procedure is well-defined for comprising the arrangement of associated words, which is optimally arranged in a line pattern test, with the sequence of isolated word pattern signal. One is suggested by Sakoe while solving the same problem which shown in the given algorithm. Implementation of the constraints for the level of edifice algorithm is accessible. It was more adaptive and effective, but more difficult. The clarified version of the algorithm was suggested by Vintsyuk.

Sakoe's algorithm moderately was complex and inconvenient for real-time applications. Level for the construction of DTW [10] algorithm was Rabiner and Myers in 1981 [90] that was associated with the given algorithm.

In speech scenario, HMM classifies with neural network. An interesting work on HMM is presented by Jesus Gomez Villardebo, Doroteo Torre Toledano, and Luis Hernandez Gomez [7]. Paper also examines the stimulus of HMM training. Formant followers mainly contain two dissimilar limitations: one is analyzed speaking and another is formant applicants acquired via commanding dissimilar restrictions, possible formants are chosen. Next constraint gets regular restraints on the formant selection procedure. New approaches join phonemic information used for orthographic transcription of the speaking signal statement. Phonemic sub division is gained by speech itself. This algorithm is discussed in detail analysis on the

performance basis gives better results by using different techniques but it is not dependent on HMMs different degrees of training.

For the betterment of performance of HMM, unlike approaches have been suggested. This approach [9] is based on multilayered tree structure to establish whole covariance matrices into a hierarchical frame. The crosswise covariance matrix skills such as, a half of leaf node at the bottom of the tree. This algorithm outstrips other covariance modeling like Semi-Tied Covariance (STC), Heteroscedastic Linear Discriminant Analysis (HLDA), and Modeling Mixtures of Inverse Covariance (MIC), direct complete covariance modeling.

In 1990's, there was a paradigm shift in the field of sequence recognition. In 1990's, research was done to rise the strength of speak recognition system. Various techniques were investigated the incompatibility on and amongst testing & training condition affected by noise, speech eccentricity, broadcast channel etc... Major techniques included MLLR [49], Parallel Model Composition (PMC) [74], Structured Maximum a Posteriori (SMAP) [75] were investigated.

In early days the idea of neural network was useful to speaker recognition structures. Artificial neural network method is a mixture of audio phonetic & sequence recognition approach. The ANN speech recognition system recycled to recognize utterances and distinguished different sound signal as the way as human being would recognize.

Even though the high recognition performance is achieved using the Neural network, it is not much successful method because of the following limitations. The first is related to the requirement of computing resources that limits the performance and second is its inability deal with run able behavior of speech. To overcome the above problems, neural networks should represent temporal relationship between acoustic events, while at the same time provides invariance under translation in time that requires precise segmentation and aligning methods. The processing of speech signals is high time consuming and not fully reliable.

The Defense Advanced Research Project Agency (DARPA) funds the project on speech understanding, which led to various technologies. The first speech understanding system was 'Hearsay I' developed by Carnegie-Mellon University (CMU). The 'Hearsay I' system was able to use semantic information to significantly

reduce number of alternatives considered by recognizer. The other system under the project was Hearsay II by CMU and HWIM (How what I mean) by BBN. The Hearsay II uses similar asynchronous procedure that pretends the component education. The efforts of CMU with Sphinx systems were to successfully integrate the arithmetical technique of HMM with the system search is the strength of earlier Harpy system.

The various application systems were developed like programmed speech document indexing and rescue systems. In 2000s, the text EARS sequencer was directed by DARPA& makes it as soon as possible for machine. In year 2000, the major focus was to enhance the recognition performance for spontaneous speech. Accuracy of results for ASR has achieved acceptable levels but these results drastically decline when it comes to recognizing spontaneous speech. To broaden the application and reliability of automatic speech recognition the performance of spontaneous speech recognition should be raised. In order to grow recognition act for natural speech, some projects were held. In Japan "Natural Speech: Amount and Processing Techno" was held, and several original techniques including elastic audio modeling, judgment boundary recognition, accent modeling, language model adaptation as well as acoustic and automatic speech summarization [11] [55] were investigated and reported.

The other focus area remaining is increase of strength of speech recognition. To enable the dialogue application having interaction level like humans that is associated with every recognized occasion with numerical so that ASR system can be used to accept these recognized events confidently [58]. These confidence measures in turn serve as a guide who can be used to provide responses and also detect significant part and reject irrelevant portion in spontaneous expressions. Further investigations have been done in combining visual and audio information for multimodal speech recognition [42].

1.11 GENERAL REVIEW OF FEATURE EXTRACTION TECHNIQUES

Now we review the approaches of features extraction techniques used in automatic speech recognition.

***Approach In Transform Domain*:** In this method speech features are taken out in the converted domain, capturing significant fact and discarding dismissed data from the speak signal. Some of the popular transform domain approach are the filter banks, Discrete Fourier Transform (DFT), Linear Predictive Coder (LPC) and transforms for orthogonal such as Discrete Cosine Transform (DCT) and Karhunen Loève transform (KLT).

Auditory spectral model includes the open simulation of the human auditory model in positions of activity in the cochlea. Most standard models are Ensemble Interval Histogram (EIH). The drawbacks are relatively poor in performance. The Ensemble Interval Histogram (EIH) [67] is very much capable of very nice recognition process.

This paper recommends the perceptual invariance to hostile the speech signal situations (microphone, noise, network falsifications, & room echoes) and the phonemic variability (because of the unsymmetrical articulator gestures) may afford a basis for vigorous speech gratitude. Author also designates the state-of-the-art for acoustic model. It pretends the outer parts of the aural fringe through the auditory bravado level. In the given broad sheet, the speech signal data is pull out through the replicated auditory nerve sackings. The author discusses introductory experimental results that ensure the human usage of such kind of assimilation, with the diverse integration rules for different time-frequency provinces depending on the phoneme-discrimination task.

Payton model [42] integrates the processing stages that relating the transformation from the auditory pressure-wave electric signal at the eardrum for different time activity in acoustic neurons. This prototypical involves the concatenation of components, for every section of the boundary. All the modules are based on circulated algorithms and on the current tentative records, except the basilar membrane i.e. presumed to be undeviating. Although the segregation of nonlinear membrane mechanics, the accurately model predicts the speech vowel formant representations in the Average Localized Synchronized Rate (ALSR) rejoinders and the saturating signal are the characteristics of the standardized average rate those responses inconspicuous.

In the paper by R. Lyon [67], he claims that speech signal analysis algorithms should be based on traditional or computational models of human audition, starting at the ears. This paper proposes models of the inner cochlea, or ear, which are expressed as place-and- time-domain signal processing operations; i.e. the models that have computational expressions for the functions of cochlea. The main parts of the models concern the mapping of mechanical vibrations into neural representation and mechanical filtering effects. Compared to other speech signal analysis techniques, this model provides a much better job of both time and frequency, which is important for robust sound analysis.

The computational good pattern suggested by S. Seneff [76] to model the human acoustic system is called mean-rate model. Then it goes to remove the very high and very low frequency modules.

***Spectral Density Methodology*:** This paper determines the functional relationship between pitch and frequency [77], so that pitch, as a perceived aspect of the tone, could be measured. Two methods (equal sense-distances and fractionation) were used to determine such a scale and they gave very good agreement. The scale unit derived is called a Mel and is defined as 1/1000th of the pitch of a tone. It was successfully used by Fant [12], Picone [35], Mermelstein and Davis [60] to remove features from speaking signal for better recognition performance.

In the paper by Picone four simple processes of signal modeling, that is spectral modeling, parametric transformation, spectral examination, and numerical modeling, are debated. Three significant trends were developed in speech sensation. First, assorted parameter sets that mix total spectral information by time-derivative, spectral data, become mutual.

In the paper by Mermelstein and Davis numerous parametric illustrations of auditory signal were matched. Every parameter set (constructed on a Mel-frequency), a rectilinear prediction cepstrum, true frequency cepstrum, a set of replication constants, and term patterns were generated by using a well-organized run able method.

The book “Acoustics and Vibration Physics” is written by Bate and Stephens [70]. They specified Mel spectrum in relations of the acoustic model.

***Signal based process*:** Given method allows them to eliminate the necessity of speech preprocessors. It usually serves a part of speech waveforms conversion addicted to frame-based speech signal facts focus on following modeling procedure. Now this signal is animated via a Gaussian foundation i.e. the time-varying in power. Established on outcome of the examination, they suggest & assess regularization planned to eliminate the effective compassion of speech signal recognition.

In joint time frequency domain method, spectrograms were the first kind of Time-Frequency Distribution (TFD) used in signal modeling. In this Paper they have applied two iterative methods for generating positive time-frequency distributions (TFDs) to speech analysis [5]. These methods provide use much of sources of information. Plosive events and formant harmonic structure are simultaneously acquired in TFDs. Slowly the time-varying formants are resolved by these TFDs, and harmonic structure is also revealed. The free sweep rate; provides a quite different result from seen with conventional speech spectrograms.

Reily and Boashash linked with TFDs /wavelets [61]. They bared the short-time Fourier transform technique that performed better than the bilinear dispersals when the window and frame period values were very enormous, as desired via the speech signal communal. He also exposed that, auto deteriorating model reflection measurements that performed consistently, typically better than all and both time-frequency cataloguing which performed the STFT for very small windows and higher time resolutions. Performances were dropped in most of the cases when quantization level is increased. By using window parameters, the result was significantly poorer than those that use much smaller windows. Whereas Rainton and Young [15] equated TFDs with clean bank energies like speech feature vectors. In particular, Wigner distribution has performed well. Despite its improved time-frequency resolution, employing the Wigner distribution (WD) as front end of speech signal recognition system actually improves recognition performance; only by explicitly re-introducing time-frequency smoothing into the Wigner distribution are recognition rates improved. Author presents a practical adaptive smoothing algorithm which attempts to match the degree of smoothing introduced into the WD with the time varying quasi-stationary speech waveform regions. The recognition performance of the resulting adaptively accurate estimator is found very comparable results to the

conventional or traditional filter-bank estimators, yet the average temporal sampling rate of the spectral vectors is reduced by around a factor of 10 in result.

***Sculpting*:** Uttered speech has a not positive value of spectral slope, because of physiological appearances of the speech creation mechanism.

The system described in this paper a huge signal raw material (CSR) system and the output acquired by using the Wall Street Journal-based database. It also can be structured to accept a length stream of input speech that is unbounded.

Heck, Murthy, and Beaufays [25], used a perceptually encouraged feature such as the Mel slope for utter re-identification i.e. terminated with telephone networks. This paper addresses the issue of independent -set text-closed speaker identification from all samples of speech signal recorded over the microphone. It emphasis on the effects of acoustic signal mismatches between two phase training and testing speech data, and concentrates on two different approaches, features extraction that are robust against channel variations in communication and transforming signal the speaker models to compensate for channel effects. In the context of feature extraction, the performance gain is founded by attaining only system features (filter bank-based cepstrum and filter bank based cepstral slope). In the context of model transformation, the fixed-target compensation algorithm has founded in a significant performance gain. It has not completely compensated for communication channel effects.

Current popular feature extraction trend is wavelet packet which is based on feature extraction technique.

It has been studied by Nikos Fakotakis, Mihalis Siafarikas, IosifMporas, and Todor Ganchev [50]. In this broadsheet, the enactment of four wavelet packet-based on speech parameterization against the traditional Fourier-based techniques that was studied by using Texas Instruments and Massachusetts Institute of Technology (TIMIT) database with Sphinx-III speech recognizer. The experimental consequences were found out the wavelet packet-based speech signal. This validates wavelet packet-based speech signal with the parameterization of the schemes as an auspicious research direction.

1.12 LITERATURE REVIEW OF SPEECH RECOGNITION FOR INDIAN LANGUAGES

There is very initial stage of systems for Indian languages and the technology is very less available. The Ist speech recognition system our country was developed for Hindi language [51-52]. The main aim or motive of this recognizer was to identify the isolated letters in Hindi. Same work has been done for the Tamil language also. In case of Malayalam, no work has been reported till, through has the start of this language mother tongue our country. The wavelets [40] have been used for speaker classification by H. A. Patil, et al. in Marathi language. They used wavelets as polynomial classifier for speaker classification.

For this work Hindi is chosen as language because it is highly spoken language in India or many of the people using this language as the communication medium for conversation between two humans being to complete their daily needs. It is very much adopted by all the communities as basic language. This language doesn't have any different phoneme and alphabet.

This has merit in case of phoneme based recognizers; In 2004 N. Udhay Kumar *et al.* [55] has done multilingual speech recognition for the specific country. This research has lot of problems in design of the hidden markov model. Design of this research is using for Tamil and Hindi languages and also for the English. Some special findings of this research are highlighted in the unique work of the above researchers. This unique work has been taking good approaches for the work and also justified the findings.

Madelaine Plauche *et al.* [51] had presented very helpful approach for the single digit and isolated digit system work. In this data collection was very unique for the different type researches in each field of the speech. The data collection has been done during the interactions between the villagers and other persons. Their work design was very much accurate, multipurpose and modifiable to the database.

Design of above work estimating the database during the process of gradual integration and initialization [14] of signal data bases.

M. Kumar *et al.* [52] have proposed continuous two techniques for speech recognition system. These were very good and helpful approaches for the continuous speech data.

R. K. Aggarwal and Mayank Dave [68] proposed an approach to implement an ASR. Feature extraction technique LPCC and HMM to speech signal acoustic designs are very highly using for Indian language. Using this approach, a prototype of the recognizer for isolated word or digits, speaker [41] dependent ASR system for Hindi is made. To generate acoustic model for any Indian language, the design of such system is a very easy way. The use of such an ASR gives the solution of problem of technology acceptance in India by bringing human-human interaction closer to human-human interaction. They have done experiments with different number of trainings, experiment with different vocabulary sizes and experiments with different noise environments.

Shivesh Ranjan [78] has implemented discrete wavelet transform for Hindi language. In his approach Discrete Wavelet Transform is explained for the language recognition. In this research coefficients of speech are computing first. After this process LPCC with discrete wavelet transform coefficients are estimating here. This method involving K means algorithm. For the above work vector quantized codebook is being used. This unique work has been giving good and helpful results.

R. Rajeswara Rao *et al.* [69] for this work isolated digits have been taken and MFCC is using to make Hidden Markov Models. Password system is also being used for the research work.

A statistical approach towards the recognition of Hindi language words is explained by Vinay Kumar [87]. In his paper he has shown how HMM is used for Hindi language alphabet recognition purpose. To prepare a model of Hindi speech author used LPC, Vector Quantization (VQ) for front end processing of speech signals. While at the back end HMM was used for recognition purpose. A noise elimination model is also used to eliminate the undesired frequency signals.

Tarun Pruthi *et al.* [86] described the real-time recognizer for Hindi language in speech field. It has some noise related modifications also that is very useful in present scenario of the world. The focus of work is the elimination /detection

method using in this way. Findings of work are very highly satisfactory level for the further work. In Swaranjali, [86] work they employed peak capturing method and having good results. By performing good process and well training, the findings of work are very good.

D.Shakina Deiv & Gaurav [17] described automatic gender identification system for Hindi speech recognition [43] [44]. This paper presents development of a gender recognition system as the preliminary work done that can be involved with the Hindi automatic speech recognition system (ASR).

(GR) Gender recognition can help in the further development of speaker-independent speech signal recognition systems. Implementation for isolated Hindi digits µp based recognizer has been explained by Ashutosh Saxena and Abhishek Singh [3]. The method that is using advanced method but feature vector of speech signal based on the zero-crossings.

1.13 SUMMARY

This chapter includes literature review with the introduction of automatic speech recognition. The motivation, application, brief introduction of all chapters and objectives of this research is also discussed. Objective of this research is to develop robust acoustic as well as visual features for Hindi speech recognition. Hindi mother tongue has been source of inspiration. It is very much adopted by all the communities as basic language. The chapter also discusses the design issues and applications of speech recognition. A brief plan of work is mentioned in all the chapters of thesis. A brief literature review of speech recognition is given in this chapter. First of all general review of development in speech recognition is explained then review in automatic speech recognition is explained. It also explains the techniques by using development of ASR in past for other languages. Finally, a brief review of research done for Indian languages is mentioned. It is observed that very less work is available in Hindi language.