Chapter 2

Literature Review

The chapter reviews the techniques used for ASR systems. It provides technological overview of the growth in this field. We explored the literature to formulate the objectives of the proposed research. Literature is analyzed for state of the art of technologies used in ASR system in relation to the problems being explored. This review covers both ends of ASR, i.e. front end (feature extraction) and back end (acoustic modeling).

2.1 Introduction

Research in the field of ASRs can be traced back to early 50's with advent of digital computing technology. It was possible through conversion of analog to digital speech information form. It was made possible using spectrogram for discovering discriminative information in feature vectors of different classes. Initially, digit recognition was investigated by bell laboratories (Davis et al. 1952). In 1960's it launched automatic segmentation of an input speech with advances in pattern matching and modeling classifier approaches. Later, it introduced mapping of segmented speech elements into its linguistic unit - phoneme, syllables and their corresponding words. In earlier aspects of constructing an ASR system, these phonetic elements were well described for their articulation and in context of its phonetic helping in generation of a sound. In last four decades, advances in pattern matching and pattern comparison were used to build representation of a speech pattern. The speech pattern was earlier represented in particular template or using statistical model such as HMM. The classification of pattern was made through four approaches: template based approach, statistical approach, neural network approach and their hybridization. In 1970's, the building ASR was only made possible using Discrete Wavelet Transform (DTW) techniques. This approach stayed for a long period of time that tried to synchronize similar segment of acoustic component in according to its reference patterns. Furthermore, in 1980's, the popularity of statistical approaches was first applied in mid of 70's (Jelinek, 1976). In late 90's e.g. for DARPA projects, various new speech reorganization systems and techniques were sprung up. These techniques with more refinements showed further improvement in accuracy. It was made possible through addition of data driven approach using language modeling technique (Beaufays et al. 2003). A survey of 30 years for design challenges, refinement and aspiration were founded by key researchers in the domain of speech technology. However, with passage of time an urge for up gradation in techniques of pattern matching were tried that goes beyond to its conventional backend approaches framework. Traditionally due to lack of computing power, the training process of the ASR system took days and weeks together. Nowadays, it has reduced to a few hours. ASR system in older times identified on basis of spoken speech (spelled speech), isolated, connected, continuous, spontaneous and conversational speech. But, these days ASR systems are being built using techniques: HMM, DTW, GMM, Deep Belief Networks (DBN), SVM, GMM, Subspace Gaussian Mixture Model (SGMM), ANN, DNN and their hybrid models used in modeling classifiers, MFCC, RASTA, PLP, MF-PLP, GFCC, PNCC and CFCC at front end and RNN, LSTM in language modeling phase. Efforts are going to improve the training and testing phases using complex models. Basic architecture of an ASR technology taken into account the feature extraction and modeling phase as an important activity for development of a successful ASR engine. These phases play a crucial role in two modes: training and testing through feature extraction methods as discussed later in Section 2.2, and the better output generation in decoding phase is presented through proper modeling approaches in Section 2.3. Section 2.4 describes the state of the art of corpus of different Indian ASR engines. The present working progress of Punjabi ASR system is represented in Section 2.5. Comparison of different Indian language ASR engines with various front end, back end and toolkit adopted are discussed in Section 2.6. Finally the research gaps, objectives of the proposed research and summary of reviewed literature are showcased in Sections 2.7 to 2.9.

2.2 Feature extraction phase

For many years feature extraction process in ASRs, has been an active area of research. Representation of compact information is a primary requirement. Efforts have been made to select the important variable subset from an input signal by filtering out noise and irrelevant variables. Different classical feature extraction approaches have been used such as Linear Prediction Coding (LPC) in 1971, Linear Prediction Cepstral Coefficient (LPCC) in 1974, MFCC in 1980, Preceptual Linear Prediction in 1990 (PLP), RASTA-PLP in 1994, and PNCC in 2010 with static, dynamic or combined feature vectors. In clean environment, MFCC technique was mostly employed which can efficiently extract known speech variations closely replicating human ears. PNCC technique was proposed by (Kim, C., & Stern, 2016) which include power-law nonlinearity after deployment of the log. CNN based raw waveform model (Variani et al. 2016) tackles the issue of learning through two cascaded stages such as filtering that is consecutively followed by pooling stage in baseline log mel model. Further it employs complex value linear projection for features replacing the CNN model. It was used in MFCC to provide substantial improvement than baseline MFCC and PLP features. A block based transformation technique was applied on baseline MFCC. It overcomes the issue of covariance matrix calculation with log energies and its full-band was evaluated for MFCC features. These features were corrupted through narrow band channel noise. Experiments were processed on National Institute of Standards & Technology (NSIT) Speaker Recognition Evaluation (SRE) dataset with GMM-UBM model (Sahidullah, M., & Saha, 2012). Analysis of two aspects of feature (MFCC and

PLP) was investigated after combining them by LDA and log linear model approach (Zolnay et al. 2005). An auditory based robust feature extraction approach was presented to decrement the string recognition error rate (in close-talking and in hands free environment) on connected digit recognition system (Li et al. 2000). A Cochlear Filter Cepstral Coefficient (CFCC) based auditory transformation was subsequently investigated (Li, Q., & Huang, 2011). It helps in analysis of the acoustic mismatch in training and testing environment for an input speech. CFCC performs better than MFCC when trained in clean and tested on noisy or mismatched conditions. Gammatone filter based robust auditory features were extracted under mismatched train and test conditions using GFCC approach. It was found to be performing better than MFCC, RASTA-PLP and CFCC feature vectors (Li, Z., & Gao, 2016). GFCC was also studied (Zhao, X., & Wang, 2013) for noise robustness and was found to outperform MFCC. Lossy inversion issues for model parameters of Mel-Frequency Cepstral (MFC) and Mel-Frequency Spectral (MFS) domain were examined for mapping them in a spectral domain. They were addressed with Parallel Cepstral and Spectral (PCS) based modeling on HMM approach (Veisi, H., & Sameti, 2013). Combinations of MFCC, PLP, VLTN and MF-PLP features were performed with Gammatone features using weighted ROVER was found to provide better results than log-linear model combination and LDA (Schluter et al. 2007). The wavelet packet based features were obtained for TIMIT speech recognition using CDHMM model (Farooq, O. & Datta, 2010).

2.2.1 Filterbank optimization

A filterbank is employed in preprocessing form in feature extraction and their reduction from input vector information. An unsupervised learning algorithm was contributed (Gautama, T., & Van Hulle, 1999) helping in development of a parameterized filter. Experiments were performed on real application dataset. A data driven analysis was performed by designing of modulation filter algorithm boosting the performance of an ASR system (Chiu, Y. H. B., & Stern, 2014). The environmental and filter distortions were further tackled by constrained optimization process for DARPA resource management and Wall Street Journal corpus on Sphinx-III toolkit. A research that efficiently controlled the issue of mismatch between train and test condition was presented by obtaining new auditory based features (Li, Q., & Huang, 2010). The invertible time frequency transformation was used for robust speaker identification. It was observed that cochelar filter bank performed better than baseline MFCC, and Modified-GFCC (MGFCC) feature vectors. A bio-inspired auditory model was evaluated that simulates the middle ear filtering with a low pass filter. Gammachirp auditory filterbank (GcFB) was tested on real word noise that outruns the classical approaches - PLP, LPC, LPCC, and MFCC (Zouhir, Y., & Ouni, 2014). An Adaptive Bands Filter Bank (ABFB) was presented (Huang et al. 2011) for flexible bandwidth and center frequencies utilization with GA for optimization of design parameters.

2.2.2 Low variance feature estimation

The feature vector output is the key component in the first step that strongly affects the later stages - modeling and knowledge representation in decoding phase. The feature vectors were computed through an estimated spectrum. The representation of a proper estimate was a crucial activity. This task was performed with windowed periodogram (Harris, 1978). The window periodogram has high variance and it is required to attain low variance. Multi-taper approaches were elegant methods for deployment of window peridogram (Riedel, K. S., & Sidorenko, 1995; Sandberg et al. 2010; Thomson, 1982). The impacts of these approaches were analyzed in speech recognition and speech enhancement applications (Hu, Y., & Loizou, 2004) The NIST 2002 and 2006, SRE dataset were demonstrated with GMM-UBM system (Kinnunen et al. 2010). The selection of an appropriate technique was a challenging task because numerous taper estimation methods were studied and tested in literature - sine taper (Riedel, K. S., & Sidorenko, 1995). MFCC based spectrum estimation using three multitaper methods were analyzed by (Alam et al. 2013) on NIST 2002, SRE dataset with PLDA models.

The experiments were performed on telephone and microphone datasets which obtained better results than conventional single taper approach. The observation was analyzed without optimizing i-vector speaker verification system. MFCC feature vectors using low variance approach was demonstrated (Alam et al. 2013) on Aurora - 2, 4 and NIST 2010 corpora for speaker recognition process. It obtained better results than baseline window method. Multitaper was employed (Kinnunen et al. 2010) for robust feature extraction using MFCC approach (in clean and real environment) due to their insentivity to each type with variations in the value of taper. The mean square cepstrum estimation approach with AR-estimator were used (Hansson-Sandsten, M., & Sandberg, 2009) for sinusoidal multiple window. The new specified estimator produced high recognition performance if it carries large spectral dynamics. Multitaper and Gammatone filter approach of feature extraction were used in robust speaker verification system. Performance improvements were demonstrated on TIMIT dataset by characteristics of low variance of multitaper approach and robustness of gammatone property on GMM-UBM modeling classifier (Meriem et al. 2017). Cepstral coefficient estimation was investigated with multi-taper based spectrum estimator where weights were optimized by Taylor expansion of a log spectrum variance for simulated speech signal (Hansson-Sandsten, 2013).

2.3 Acoustic modeling classifier

In Acoustic Modeling (AM) phase, the input audio signals are represented for discrimination of classes' information of baseline speech units. The input signals are matched with parameter sets of classified information**.** Feature vectors are modeled using classification approaches - DNN-HMM, SGMM-HMM, GMM-HMM, and baseline HMM or DTW combined with various optimization approach, most commonly used in ASR system framework. The process of investigation of ANN and HMM approaches were started since 1990s. In recent years, different analysis has been investigated with release of hybrid or end to end DNN approaches. These approaches were combined with baseline HMM technique to build a successful ASR system through process of hybridization or using tandem approaches (Yu, D., & Deng, 2015).

2.3.1 Hidden markov model

HMM was a defined approach for speech recognition application (Deng, L., & O'Shaughnessy, 2003; Rabiner, L. R., & Juang, 1993). Its parameters contain the characteristics of self learning from its training data. Earlier Maximum likelihood approach was employed to learn the parameters but later efforts have been applied to improve them (Deng et al. 2005). To make possibility of estimation problem a more solvable competent learning algorithm was presented - Expectation-Maximization (EM) (Neal, R. M., & Hinton, 1998). On the other hand, various generative models were explored (such as CDHMM) using Baum-Welch algorithm. It was focused upon EM technique. To enhance the efficiency of speech recognition system, various paradigms like neural network techniques, discriminative and connectionist approaches with HMM were presented (Aggarwal, R. K., & Dave, 2011). In the recent years, researchers tried to improve HMM parameters or hybridize HMM using DNN, RNN and TDNN training techniques. For refinement and advancement of HMM numerous alterations and extensions were discussed.

2.3.1.1 Optimization of HMM approach

Hybridization of HMM is the most used method for improving of the performance of a large vocabulary speech recognition system. In a input speech signal, Markov property of HMM was found to be helpful in modeling of its temporal. HMM was naturally trained using Baum-Welch algorithm through maximization of its global model likelihood probabilities. It was employed on gradient search approach for restimation of input speech feature vectors. It was made possible using ANN approach. It avoids the problem of dynamic control in a non-linear system. It helps in query learning by Time

Delay Neural Network (TDNN) approach for a speech recognition system (Hwang, J. N., & Li, 1992). Model space approach of optimization was also used with gradient method for HMM inversion. It reduced the discrepancy in joint model and its feature space. Joint model was performed better than its single space. It was combined earlier with model space or with feature space (Moon, S. & Hwang, 1997). Gradient was mainly focused on local optimum and tried to overcome it using Tabu Search (TB). TB was integrated with HMM SI continuous speech recognition system using multi-Gaussian CD model (Thatphithakkul, N., & Kanokphara, 2004). TB was also applied on HMM for optimization of model structure. It performs better than a conventional method (Chen et al. 2004). A comparative study of various model architecture (using hybrid HMM) was examined on TIMIT phone recognition (Johansen, 2007). It was employed by global discriminative training methods that outrun slightly better than its baseline HMM approach. Continuous HMM (CHMM) with GA was proposed (Takara, T., Iha, Y., & Nagayama, 1998) helping in generation of optimal structure of HMM. It was more adequate than discrete-HMM. The efficient HMM models were also obtained using GA and Baum Welch approaches which help in finding of better number of states (Kwong et al. 2001). Baum Welch approach optimization was performed using Chaos Algorithm (CA) for continuous speech recognition. It helps in retrieving global or sub optimal solution for initial values of HMM parameters. CA was performed efficiently in comparison to heuristic approaches - PSO and GA optimized BW (Cheshomi et al. 2010). An empirical approach was presented (Aggarwal, R. K., & Dave, 2015) to produce the optimum value of a Gaussian mixture as well as degree of state tying for small corpus. It uses PLP and RASTA techniques at front end and HMM with Gaussian mixture at back end. It overcomes the issue of overfitting of data and reduces computational overhead with mixture tying characteristics. It was achieved by varying mean and variance while keeping static value of mixture components. The training of HMM were implemented (Yan et al. 2006) using Gaussian kernels allocation with nonuniformly among its states. It helps in optimization of its discriminative training conditions. The kernel digit approach was investigated on TIDIGITS speaker dependent connected dataset. The demonstration of HMM parameters with different feature sets on each state was studied in (Baggenstoss, 2001). It produced an optimal classifier to capture large information from a raw input signal. HMM topology method was generated using various combinations of objective function with addition of topology construction methods. Selection of topology was deployed using GA for Thai phone recognition methods (Bhuriyakorn et al. 2008). Evolutionary algorithm was implemented (Figielska, E., & Kasprzak, 2008) for training of HMM model. It analyzed the computational complexity for 20 words. HMM faced the issues of convergence to a local optimal solution. It was not able to fit with sparsely connected time-series data information. This issue was resolved using Recurrent Hidden Markov Model (RHMM) (Dong et al. 2011). It was trained with PSO through adjustment of distributed extremum. Training of HMM model was conducted using Maximum Model Distance (MMD) or GA algorithm on T146 word alphabet dataset (Hong, Q. Y., & Kwong, 2003). GA resulted into higher computational complexity but with better recognition output. A new application of PSO based HMM training was adopted (Macaš et al. 2006) to traverse the optimal value of model parameters using stochastic characteristics. It helped in handling constraints using three different options. Its performance analysis was compared with baseline BW algorithm. Furthermore, training process was improved with optimized model parameters values by using Baum Welch Genetic Algorithm (BGA) (Mehla, R. & Aggarwal, 2013). A frame based cross entropy training was employed in HMM/NN acoustic model with sequence classification criteria for English broadcast news (Kingsbury, 2009). NN-HMM worked on speaker normalization (using local filtering and max-pooling) achieved higher speech recognition performance (Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, 2012). Parallel Genetic Algorithm (PGA) (Kwong, S. & Chau, 1997) outperformed the heuristic approach.

Model parameters were processed with PGA for HMM training. GA based training of HMM model with MCE approach was presented as a re-estimation paradigm (Kwong et al. 2002). Feature subset as well as model parameters were optimized using GA for discrimination among HMM based classifier (Li et al. 2010). Optimal Discriminative Training (ODT) was also used to continuous mixture density HMM. It improved the performance upto large extent in noisy environment of SI speech recognition system (Mizuta, S., & Nakajima, 1992). To refine the model parameters of HMM, optimization algorithms were traversed and helped in performance improvement of baseline HMM system.

2.3.2 Deep neural network approach and its optimization

In recent years, DNN has attained maximum attention with alleviation of hardware and refinement of machine learning algorithms. DNNs were integrated earlier with HMM. Modeling sequence of acoustic observation performed better than GMM-HMM (Deng et al 2013). The researchers employ baseline HMM or GMM-HMM for modeling of its state information (Liporace, 1982). DNN issues overcame the issue of GMM-HMM through demonstration of a high quality in modeling of data space nonlinearities (Hinton et al. 2012). It resulted into popularity of DNN-HMM model for modeling classifiers in an ASR system. Their exists a few places where vocabulary size is small, so it was implemented using GMM-HMM models. It occurred due to availability of limited amount of training corpus that helps in learning and modeling of classes. DNN were combined with Context Dependent Deep Neural Network HMM (CD-HMM) to explore GMM processed features. It was employed in baseline bottleneck feature vectors that help in generation of better recognition performance (Tomashenko et al. 2016). CD-DNN-HMM approach was also examined (Lee et al. 2015) to improve large vocabulary continuous speech recognition using two feature engineering aspects. The proposed feature approaches outperformed directly using baseline MFCC (on GMM) and mel filterbank output (on DNN). DNN failed to perform on small or limited vocabulary size (Lasserre et al. 2006). A small amount of data was taken into consideration for training of neural networks. It was proposed on posterior probabilities (Mao et al. 2006). Studies focused later on generation scheme for pseudo-sample. It uses GMM for production of artificial samples. It was further employed as an input to DNN for learning of its generalized structure. The generative information were combined with real training data to reduce recognition performance (Ghaffarzadegan, S., Bořil, H., & Hansen, 2017). DNN was also processed through numerous adaption models - linear transformation using Linear Input Network (LIN) adaptation (Gemello et al. 2006; Li, B., & Sim, 2010), feature space Discriminative Linear Regression (fDLR) (Seide et al. 2011; Yao et al. 2012b). These models helped in stimulation of a number of hidden layers in Linear Hidden Network (LIH) (Gemello et al. 2006) adaptation and in softmax layer**.** DNN methods developed till date adopted the benefits of robustness of GMM (Seide et al. 2011). GMM was combined with DNN for transformation purpose**.** By using GMM feature in fMLLR method, it act as an input for training of a DNN model (Parthasarathi et al. 2015). In decoding phase, the state level information was used through likelihood scores on same fMLLR features. It helped in temporal changing of weight regression and utilization of GMM-derived feature information (Liu, S., & Sim, 2014; Tomashenko, N., & Khokhlov, 2015). The investigation of HMM free architectures were also presented (Maas et al. 2017) with selection of DNN parameter like number of layers in model size, along with its architecture and through its training details on Switchboard corpus of around 2100 hours of training dataset. The outline of Table 2.1 represents the neural network based ASR engine.

Table 2.1: Outline of NN based ASR engine

Language	**Corpus Size**	**Modeling Classifier**	**Performance Analysis**	**Reference**
Russian	16,350 spoken by 50 Russian Speakers	GMM-HMM, DNN-HMM	20% WER reduction as GMM-HMM system	(Kipyatkova, I., & Karpov, 2016)

Malayalam	108 Sentences with 540 words and 3060 phonemes	ANN-HMM	86.67% word and 66.67% sentence recognition rate	(Mohamed, A., & Nair, 2012)
Bengali	SHRUTI Bengali Continuous ASR Speech Corpus 1050 voice samples	HMM-GMM, DNN	40.19 %	(Reza et al. 2017)
Japanese	4000 sentences	DNN	Mean squared error value is significantly reduced	(Takamichi, 2017)
English	CSLU Foreign Accented English (FAE) corpus.	DNN	Test Accuracy 47.9%	(Siddhant et al. 2017)
Romanian	200k words	DNN	Relative WER improvement 18% to 23%	(Georgescu et al. 2017)
English	TIMIT and WSJ with noise and reverberation	DNN	12.9% WER	(Ravanelli et al. 2016)
Russian	40M words with vocabulary size of about lOOK words	RNN	Relative word error rate reduction of 14% with respect to baseline 3-gram model	(Kipyatkova, I., & Karpov, 2015)
Russian	1 million word type vocabulary	DNN	Reduction of WER , 4.4% relative, compared to a robust n–gram baseline model	(Renshaw, D., & Hall, 2015)
Mandarin	RASC863 corpus	RNN	77.23% WA	(Ni et al. 2016)
Mandarin	RASC863 and CASIA regional accent speech corpus of 260 hours	RNN	5.8% WER	(Yi et al. 2016)
Bengali	CRBLP and CDAC corpus	DNN	88.44% WA	(Bhowmik et al. 2015)
Bengali	2000 words	RNN	WER 13.2% and PER 28.7%	(Nahid et al. 2017)
Arabic	TIMIT datasets	CNN, DNN	80.75%, 72.0% WA	(Rajagede, R. A., & Dewa, 2017)
Arabic	1200 hour	DNN	18.3% WER	(AlHanai et al. 2016)
Hindi	1 hour of transcribed Hindi dataset and 15 hours of Wall Street Journal (WSJ) English data	DNN	16% phone recognition accuracy increase	(Dey et al. 2014)
English	NIST Open KWS’13, REPERE, AMI Meeting Corpus, NIST Open SAD’15	DNN	5.76% WER	(Gelly, G., & Gauvain, 2018)
Mandarin	AMI, HKUST, GALE, and MGB	DNN	48.3%, 31.36%, 22.33%, 22.18%	(Hsu et al. 2016)

English	WSJCAM0 Cambridge Read News REVER corpus	DNN-HMM	25%, 27%, 26% WER on Near, Far, Avg. Real data	(Huemmer et al. 2016)
Mandarin-English	LDC2015S04	Multi-task Learning MLT-DNN	MER (Mixed error rate) reduced by 12.49%	(Song, et al. 2017)
English	U.S. English Broadcast News (BN) transcription task	Multi-basis DNN	9.8% WER	(Karanasou et al. 2017)

2.4 Analysis of Indian ASR corpus

A robust spoken language processing system such as automatic speech recognition depends heavily on availability of an authentic speech corpus. Only a few languages in India have associated speech recognition engines viz in Hindi and English. ASRs in Indian languages are still in their nascent stage of research and getting increased attention nowadays. Only a few languages till date are able to assemble speech corpus in their native resource. The major issues occurred was due to existence of dialectal and tone difference within the region of different languages. It gives birth to an issue of collection of a huge corpus in particular language. Building speech corpora for an Indian language is a difficult task. Statistical approaches used for modeling of an ASR engine depend upon large amount of training corpus that helps in recognition of uttered instances. Thus it is important first to have database that comprises of all characteristics of users who can speak data in realistic environment. A speech dataset is of two types: dataset collected in a particular task domain and a general purpose dataset. Previously (Chourasia et al. 2005) a Hindi language corpus was built with 50,000 Devanagari script to develop an ASR system. The speech corpus of different language ideas were designed and evaluated for Marathi by TIFR and IIT Bombay (Godambe, T., & Samudravijaya, 2011), Hindi language travel domain dataset by C-DAC Noida (Arora et al. 2010). Telgu language dataset for Mandi information system by IIIT Hyderabad, English, Hindi and Telgu dataset for travel and emergency services were collected by IIT Hyderabad (Mantena et al. 2011). General purpose corpus of continuous Hindi sentences were developed by TIFR and C-DAC Noida (Samudravijaya et al. 2000). Another general purpose corpus of Telgu, Hindi, Tamil, and Kannada were prepared by IIT Kharagpur (Rao, 2011). Hindi, Indian English corpus were developed by KIIT, Bhubaneswar and supported by Nokia research centre China (Agrawal et al. 2012). These corpus have been studied to analyze the attempt used for development of an ASR system in Indian languages. Three language optimal databases were constructed for Tamil, Telgu and Marathi languages serving as a catalyst for research activities. Two methods were adopted to collect the corpus for a landline and mobile dataset (Kumar et al. 2005) tested on Sphinx-II toolkit. The corpus was collected in coordination of IIT Hyderabad and HP Labs, India. Indian English and Hindi language corpus was collected (Agrawal et al. 2012) to involve the versatility by mobile communication environment of 100 speakers. Constructed database was employed in mobile speech recognition services. A Punjabi speech corpus of Malwai dialect was collected from 50 speakers that help in development of Punjabi Speech synthesis system. The recorded dataset was labeled phonemically to obtain phonemic and its sub-phonemic information (Bansal et al. 2015). Issues in construction of speech corpus was observed and studied (Kiruthiga, S., & Krishnamoorthy, 2012) for Indian languages. The issue in unit selection speech synthesis of Hindi and Telgu voices were presented (Prahallad et al. 2003).

2.5 State of the art of research in Punjabi speech recognition

During 1990s, ASR systems were developed by IBM, Dragon, BBN, Cambridge, Philips, MIT and LIMSI in one particular language (English) which was successfully converted to other languages (German, French and Chinese). In development of an ASR system, researchers aimed to adapt, produce spoken resources and acoustic models for under-resourced languages. But in today’s technical world, languages are getting extinct at an alarming pace (Crystal, 2000). Indian languages such as Punjabi is declared as an under resource language due to unavailability: stable orthography, electronic resources for speech, language (small amount of transcribed corpora of speech, few length of pronunciation dictionaries and a very few vocabulary list), and linguistic experts. A small amount of research is being carried on isolated, connected or continuous corpus using HTK and Sphinx toolkits. It was declared as an under-resourced language even through being used by 105 million around world. The lack of resources in Punjabi language was overcome through pioneering speech corpus collection methods (Gelas et al. 2011), along with innovative front end and acoustic model approaches. Moreover, the issue arises were due to dialectal characteristics, impossibility of native speaker with knowledge of technical expertise in ASR technology and tonal affects in the language.

In construction of a Punjabi-ASR system, transcribed spoken utterances from native speakers, massive information in pronunciation dictionaries and huge amount of text transcribed corpus helps in effective training of a statistical language model. Collection and distribution of a large amount of speech and text corpora was made possible through Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), and AppenButlerHill, Speech ocean, CIIL and TDIL. These datasets helped in building of an ASR engine. Till date, no standard speech and text corpus were available in Punjabi language. All the research activities covered by the researchers were presented in their own self created corpus. Very few researchers carried their research on standard speech corpora. The speech corpus and its transcription were employed as a base for creation of an ASR system.

Punjabi language has a place in Indo-European family. It is considered as tonal in nature. As found in 2001 statistics of India, an aggregate of 105 million natural speakers of Punjabi language made it tenth most uttered language in the world. It incorporates lexical tone at word horizontal and phonetic aspect of tone was available in the locale of pitch. It involves part of important features like 5 tonal consonants and, inborn dialectal peculiarity. It categorized the language into 4 important dialects that were build upon its articulation marking and origin of the speaker (such as Majha, Doaba, Powadh and

Malwa) spoken in particular region of Indian sector of Punjab. The tonal and in addition regional particular of Punjabi dialect assumes a vital part in production of its speech corpus. Indian Punjabi is written in Gurumukhi script from left to right. It is still investigated as a resource scared language due to inadequacy of resources. Today a number of researches have been focused on text analysis segment and only a few exploration have been done on framework of an ASR system. In case of Punjabi language such an initiative was taken by (Kumar, R., & Singh, 2011) who worked on speaker dependent isolated words created for small vocabulary with the help of DTW approach. (Dua et al. 2012a) constructed a speaker dependent as well as speaker independent 115 isolated lexicon Punjabi ASR system that employed 8 speakers with HMM technique for creation of its acoustic model. They added extension in their research work by connected lexicons of previous 115 words on similar front and back end approaches (Dua et al. 2012b). (Ghai, W., & Singh, 2013) reported a continuous automatic speech recognition system using a baseline 100 sentences whose repetition was generated by 9 speakers with a triphone modeling unit. (Lata, S., & Arora, 2013) examined the impact of /h/ sound with the help of Praat and Matlab software on a particular dialect of Punjabi language such as malwa dataset. (Mittal, S., & Sharma, 2014) investigated three mechanism of data processing specifically for read, lecture and conversation utterances and it was analyzed with the help of continuous density HMM on HTK toolkit. (Singh et al. 2015) has explained five Punjabi tonemes set up on their position and observed them using 150 exclusive lexicons gathered from 10 speakers. Individually a single dialect corpus has been handled mostly by investigators that help in generation of an ASR system. All the specified aspects motivated us in performing research to collect native speaker’s data with basic technical knowledge on ethical rules and to implement them on refined front end and back end approaches.

2.6 Comparison of different Indian ASR engine

The development of successful ASR engine depends upon a number of factors. The selection of a particular type of feature may affect the training and testing phase of the system. These feature vectors depend upon various conditions. The discriminate feature needs were classified to an appropriate classifier. The selection of particular type of classifier also plays a crucial role. Each classifier has its own limitations and functionalities. Performance recognition of the testing phase is affected due to training through particular feature and their observation stores in a different classifier. The training of system is focused mainly upon its corpus. The size and methods of speech corpus collection may drastically affect the system output. This subsection represents the recognition performance obtained through different combination of front and back end approaches in Table 2.2.

Table 2.2: Brief overview of Indian ASR engine

Language Type	Modeling Technique	Recognition Rate
Hindi (Kumar, K., & Aggarwal, 2011)	HMM	94.63% WA
Hindi, TI-20 (Mittal, T. & Sharma, 2016)	SVM with PPO & Hooke-Jeeves	81.50% WA, 92%WA
Hindi (Mittal, T., & Sharma, 2016)	SVM, binary PSO-SVM, binary PSO & Hooke-Jeeve-SVM, binary PPO-SVM	83.7%, 89.2%, 91.1%, 90.0% WA
Hindi (Kumar et al. 2012)	HMM	87.01% WA
Hindi (T. Kuamr et al. 2014)	GMM-HMM	97.04% WA
Hindi (A. Kuamr et al. 2014)	GMM-HMM	95.40% WA
Hindi (Dua et al. 2017)	GMM-HMM, MMIE, MPE	31.14%, 27.4%, 25.9% WER
Hindi, Sanskrit, Punjabi and Telugu (Ranjan et al. 2010)	ANN	83.29% with back propagation algorithm and 92.78% using clustering algorithm
Assamese (Dutta, K., & Sarma, 2012)	RNN	90.47% WA
Tamil and Telugu (Renjith, S., & Manju, 2017)	KNN / ANN	76.41% WA
Assamese (Bharali, S. S., & Kalita, 2015)	HMM	80% WA for MFCC and 95% WA for LPSEPSTRA
Kannada (Thalengala, A., & Shama, 2016)	HMM	74.35% WA for triphone
Bangla (Hasnat et al. 2007)	HMM, SVM	Isolated for speaker independent-70% and

		continuous speaker independent-60%
Tamil (Radha, 2012)	HMM	88% WA
Kannada (Hegde et al. 2012)	SVM	79% WA
Hindi (Kumar, et al. 2004)	HMM	85.46% WA
Punjabi (Kumar, Y., & Singh, 2017)	HMM	96.87% WA , 98.71% sentence accuracy
Hindi (Dey et al. 2014)	DNN	59.90% WA
Hindi (Pandey et al. 2017)	DNN	10.63% WER
Hindi (Mandal et al. 2015)	DNN	82% WA
Hindi (Dua et a. 2018)	HMM+GMM, HMM+MMI, HMM+MPE	86.9% WA (Clean), 86.2% WA (noisy)
Mizo(Dey et al. 2018)	SGMM, DNN	10.3% PER, 23.2% PER

2.7 Research gap

1) Major studies in Punjabi Speech have focused on isolated and connected speech corpus. The area of continuous and spontaneous speech recognition is still at the initial stage of research (Dua et al. 2012a, 2012b; Ghai, W., & Singh, 2013; Kumar, Y., & Singh, 2017; Kumar, 2010; Mittal, S., & Sharma, 2014).
2) The only standardized speech corpus available officially doesn't include regional characteristics of Punjabi Language.
3) A few techniques for feature extraction and acoustic model generation have been implemented in existing studies. On other hand, with advance acoustic research many hybrid techniques have been proposed. These proposed approaches have been implemented further on Punjabi speech dataset (Abdelaziz, A. H., & Kolossa, 2016; Bharali, S. S., & Kalita, 2015; Cheng, Chen, & Metallinou, 2015; Hinton et al., 2012; Imseng et al. 2012; Imseng et al. 2011; Jiang, 2010; Kurian, 2015; Lee et al. 2015; Mehla & Aggarwal, 2013; Veisi & Sameti, 2013).
4) A few ASR systems are available that used large Punjabi speech corpus (Ghai, W., & Singh, 2013; Guglani, J., & Mishra, 2018; Kumar, Y., & Singh, 2017; Mittal,

S., & Sharma, 2014).

2.8 Objectives of the research

The objectives of the proposed research work are structured as follows:

1) To develop a Punjabi speech corpus.
2) To propose an effective technique that optimizes the acoustic features selection and classification.
3) To develop an effective automatic speech recognition system using proposed technique.
4) To analyze and compare the performance of the proposed system with traditional automatic speech recognition system.

As an output, this thesis proposes a novel & hybrid method for feature extraction and helps in refinement of feature vectors used in classification approaches. As a result, it also increases the performance of large vocabulary ASR system for training and testing.

2.9 Summary

Processing and building of a real life application in particular ASR engine is a big challenge and is a rigorous task. Therefore, few efforts have been made in front end stage that helps in improvement of the performance of Punjabi ASR system with new or refined feature vectors. A modeling approach framed by combination of HMM with hybrid neural and optimization approaches were explored to improve the system. This chapter briefly reviewed the work of various researchers who implemented different front and back end approaches individually or through refinement of their feature vectors, and hybridization of approaches. Very few researchers have worked in Punjabi language as compared to a massive population who speak and use this language. A comprehensive review of Punjabi speech recognition system was presented. This is also because of non-availability of sufficient Punjabi corpus for speech processing. In further subsections, demonstration of GMM framework through adaptation on DNN-HMM model is presented. The feature vectors derived from GMM were further employed in training of DNN models using GMM based adaptation techniques on hybrid Punjabi datasets. Therefore some of the reviewed approaches were combined and implemented on large vocabulary Punjabi corpus. Finally the review of Indian speech corpus and comparative study of various Indian language ASR systems was reported.