LITERATURE SURVEY

Literature survey briefs out the work done by various researchers in the field of speech recognition. Different features extraction techniques to extract the features from speech signals and different pattern classifier and modeling techniques to design the speech/spoken word recognition and speaker identification/verification systems are outlined. It also includes the designing of ASR systems and analyzes the performance parameters in different noisy environments.

K. H. Davis et al. (1952) designed a spoken digit recognition circuit to deal with 10 digit series 0,1,2,3,4,5,6,7,8,9 when spoken by a single talker [1]. They first analyzed the spoken digit in a manner to produce formant1 versus formant2, and then the plot of unknown digit signal distribution is compared with the plots of reference sounds using pattern matching network and most likely digit are indicated as visual information.

Franklin S. Cooper et al. (1952) conducted different experiments on the perception of synthetic speech sounds [2]. They used spectrogram to represent and to create speech sounds and pattern playback machine to convert the spectrogram pictures to speech sounds.

Bruce T. Lowerre (1976) developed The HARPY Speech Recognition System by using two earlier speech systems namely Hearsay-1 and Dragon systems developed at Carnegie-Mellon university [3]. High speed HARPY system is designed by combining the best features of both the systems mentioned above and with additional heuristics. Harpy system contains a mathematical tractable model and is designed to generate the lexicon and templates needed. The system performance was analyzed with the test data consist of speech sounds collected from four different speakers contain the same 20 sentences and 110 words.
Lawrence. R. Rabiner et al. (1979) described a speaker-independent isolated word recognition system which is based on the use of multiple templates for each word in the vocabulary [5]. The word templates are obtained from a statistical clustering analysis of a large database consisting of 100 replications of each word i.e. once by each of 100 talkers. The recognition system accepts telephone quality speech as input and works based on LPC analysis of the unknown word, dynamic time warping of each reference template to the unknown word (using the LPC distance measure), and the application of a K-nearest neighbor (KNN) decision rule. Results for several test sets of data are presented and showed comparable error rates and obtained better than speaker-trained isolated word recognition systems.

François Grosjean (1980) analyzed spoken word recognition processes and the gating paradigm [6]. Isolated words with varying lengths (duration of the syllables) and frequency (low and high) were subjected for the recognition processes. The study focused on two different experiments. They presented an experimental paradigm which was used to study the on-line processing of spoken words and extended this knowledge to spoken word recognition process. They concluded that the gating paradigm would be useful in determining the minimum amount of information acoustic-phonetic, syntactic, semantic, and pragmatic needed for the analysis of language during on-line processing.

Richard P. Lippmann (1988) submitted a paper on neural network classifiers for Speech Recognition [7]. The subject of neural networks mimics the biological nervous system of human beings. They mentioned that the difficult problems of speech recognition systems can be addressed using neural networks. They compared the neural network classifiers with the conventional classifier algorithms in classifying isolated words, vowels and spoken digits and elaborated different supervised and unsupervised neural networks pattern classifiers. They implemented neural nets and feature maps in a very large-scale integration (VLSI) device. They showed that the neural net classifiers which are using energy frames performed better than Gaussian classifiers for a digit classification problem and a Three-layer perceptrons also performed well for a vowel classification task.
Alexander Waibel et al. (1989) presented a paper on Phoneme Recognition using Time-delay neural networks [9]. For the recognition experiments, they have chosen phonemes 'B','D','G' with varied phonetic context. They have analyzed the results by comparing with HMM models. The three layered Time Delay Neural Networks approach achieved averaged result of 98.5% success rate over HMM model which achieved 93.7% from three different speakers with 1946 testing tokens.

B. H. Juang and L. R. Rabiner (1991) published their paper on Hidden Markov Models for Speech Recognition [11]. They reviewed HMM models and emphasized that these statistical framework models are mathematically precise and model parameters can be estimated with limited training set data. It also showed that the HMM models are so flexible and ease in applying to speech recognition systems in terms of training data set size, type, and architecture of the models to words/phonemes/sounds etc. They summarized that HMM based speech recognition systems have shown capable in achieving recognition rates of more than 95% word accuracy in certain speaker-independent tasks with vocabularies on the order of 1,000 words.

X. D. Huang (1992) demonstrated the powerful statistical tool Hidden Markov Models (HMMs) for the automatic speech recognition [12]. He proposed that the robustness in speech recognition can be improved by using semi continuous hidden markov models (SCHMMs) rather than standard discrete and continuous HMMs. The difference between discrete and continuous HMMs is based on the source of probability distribution functions, the discrete HMM designed based on histogram oriented nonparametric characterization and the continuous HMMs designed based on the parametric probability distribution function kernels. SCHMMs have the ability to maintain large mixture probability density functions. Two different varieties of SCHMMs along with standard discrete and continuous HMMs have been experimented on speaker dependent Phoneme classification and showed that the phoneme classification accuracy was significantly improved using SCHMMs.

Douglas A. Reynolds and Richard C. Rose (1995) introduced Text-Independent Speaker Identification using Gaussian Mixture Speaker Models [15]. Experiments have been evaluated on a 49 speaker conversational telephone speech database using Gaussian
mixture speaker models. They achieved 96.8% of identification accuracy using 5 seconds clean speech and 80.8% accuracy using 15 seconds telephone speech utterances. They also examined different algorithmic issues like initialization, variance limiting and order selection of speaker models etc. They have compared identification accuracy of the proposed Gaussian mixture models with the other modeling techniques such as Gaussian Classifier (GC), Vector Quantization code book (VQ) and Radial Basis functions.

Joe Tebelskis (1995) contributed to the research work on Speech Recognition using Neural Networks [16]. Artificial neural networks have been examined in large vocabulary, speaker independent, continuous speech recognition system. Currently, Most of the speech recognition systems are designed using Hidden Markov models (HMMs) which is a statistical framework, HMMs designed by assuming the states are hidden which may limit the performance of HMMs. These assumptions are eliminated in neural networks and different complex function also can be learned using neural networks. They designed Speech Recognition Systems using hybrid model of neural networks and Hidden Markov Models. They showed hybrid NN-HMM produces better results over normal and pure HMM system in terms of accuracy, context sensitivity.

Douglas A. Reynolds (1995) experimented on automatic speaker recognition using Gaussian Mixture Speaker Models [17]. He demonstrated that the information in speech conveys in four different levels, the first level is the words or sentences being spoken; secondly the information about the speaker, thirdly the language and the topic of conversation and the fourth level is speaker recognition. Speaker recognition classified into two tasks, speaker identification and speaker verification. In identification, the aim is to determine which voice in a known group of voices best matches the speaker. In verification, the aim is to authenticate a person's claimed identity. They used statistical speaker-modeling technique to represents the speech sounds of a speaker voice, then speaker models were built to recognize the speaker that are computationally inexpensive and capable of recognizing a speaker regardless of what is being said. Experiments conducted on four speech databases TIMIT, NTIMIT, Switchboard, and YOHO. They achieved identification accuracy of 99.5% for the complete 630-speaker population of TIMIT database and NTIMIT accuracy dropped to 60.7% for the same 630-population identification task. For verification, EERs for TIMIT and NIMIT were 0.24%, and 7.19% respectively.
B.A. Mellor et al. (1996) investigated the automatic speech recognition as an interface between human beings and multi-input devices like computers [18]. The interface between human beings to machine interactions (HMI) can be carried out through different modes rather than the manual entry and visual output modes. Different interface methods such as keypad, Trackball and combination of both, and the proposed ASR method were experimented based on the time taken for the tasks. ASR providing a speech mode of input and was found subjectively popular amongst the trial subjects but provided the lowest actual interface performance. They showed that the trackball in conjunction with ASR provided higher performance than for speech input alone.

Rathinavelu Chengalvarayan and Li Deng (1997) proposed HMM-Based Speech Recognition using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features [19]. They investigated speech recognition systems by using hidden Markov model-based (HMM) as back-end classification techniques. They developed and evaluated two versions of the optimum-transformed HMMs THMM-1 and THMM-2. THMM-1 performs state-dependent linear transformation on the mel-warped log channel energies (static) in a way that is independent of the successive frames i.e., no optimization of dynamic feature parameters and THMM-2 generalizes the THMM-1 and includes a jointly optimal transformation so as to arrive at the static and dynamic parameters together. Mel-warped discrete Fourier transform (DFT) feature interactions of front-end feature extraction showed better results over the mel-frequency Cepstral coefficients (MFCC’s) and obtained error rate reduction of 15% on a standard 39-class TIMIT phone classification task, in comparison with the conventional MCE-trained HMM using MFCC’s.

Biing-Hwang Juang et al. (1997) discussed the problem of training the speech recognition system using the classical Bayes decision theory [20]. They showed that the traditional methods relying on distribution estimation are sub optimal when the assumed distribution form is not the correct one. They compared the two different methods in the context of hidden Markov modeling for speech recognition. A new MCE approach based on learning for discrimination was discussed and the experimental results showed the minimum classification error (MCE) method performs better over the distribution estimation method. MCE method provides a significant reduction of recognition error rate.
Sahar E. Bou-Ghazale and John H. L. Hansen (2000) compared the speech recognition performance between traditional and the proposed features under stress [22]. The performance accuracy of speech recognition algorithms degrade in the adverse environment where the speaker is under stress or in emotion. In this study they have evaluated and compared the recognition results with the proposed frequency partitioning methods such as one-sided autocorrelation linear prediction (OSALPC) and Cepstral-based OSALPC features to the traditional features such as linear prediction coefficients (LPC), LPC-based Cepstral (LPCC) parameters, and Mel-frequency Cepstral (MFCC) parameters for stressed speech recognition. They focused on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. Different stressed speaking styles are considered for their study. They showed that the proposed linear prediction power spectrum is more immune than FFT to stress as well as to a combination of a noisy and stressful environment. They also showed that the alternate filter bank frequency partitions are more effective for recognition of speech under both simulated and actual stressed conditions.

Douglas A. Reynolds et al. (2000) have described the important elements of Gaussian mixture model (GMM)-based speaker verification system used in several NIST speakers at MIT Lincoln Laboratory [23]. The system is built using simple and effective GMMs for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM. As part of front end processing, firstly the speech signals are segmented into frames of 20-ms window progressing at a 10-ms frame rate, then speech activity detector is used to discard silence-noise frames, next Mel-scale Cepstral feature vectors are extracted from the speech frames. Finally, the feature vectors are normalized and linear channel convolutional effects have been removed. The UBM is a large GMM trained model to represent the speaker-independent distribution of speech features. They proved that the GMM-UBM system is very effective for task of speaker recognition.

Jaakko Hollmen et al. (2000) proposed a learning vector quantization algorithm for probabilistic models and suggested to extend the same for speech recognition [24]. They derived a discriminative training procedure based on Learning Vector Quantization (LVQ) where the codebook is expressed in terms of probabilistic models, The models
implicitly define a discrimination function in the input data space through maximum likelihood search. They proved that the algorithm in the fraud detection domain can classify mobile phone subscribers to normal and fraudulent users.

Antanas Lipeika et al. (2002) developed a Speech Recognition System based on Dynamic Time Warping (DTW) for Isolated Words [26]. They used Linear predictive coding (LPC) parameters for feature extraction and Vector quantization technique for reference templates. The performance of the proposed system was evaluated using 12 words of Lithuanian language pronounced ten times by ten speakers. Twelve Lithuanian words, including digits 0 to 9 pronounced by 10 speakers were used for recognition. Each word was pronounced ten times by each male speaker in noise-free conditions, yielding 120 utterances per speaker, total 1200 utterances collected with sampled rate of 11025 Hz. Best result were obtained using vector quantization in speaker independent mode on splitting a cluster with largest average distortions into two clusters and codebook size was chosen 32. Computation amount was significantly reduced on account of slightly increased error rate.

Vincent Wan (2003) submitted his research work on verification of speaker using Support vector machines [27]. He developed the techniques required for SVMs to work well on speaker verification using sequence discrimination approach. By representing the entire sequence as a single vector, the SVM can discriminate between whole sequences directly. Experimentally, he showed that a support vector machine combined with score-space kernel and spherical normalization can outperform on the PloyVar speaker verification database.

Qifeng Zhu and Abeer Alwan (2003) proposed analysis based Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise [28]. They designed algorithms to extract acoustic features on the noise robust part of speech spectra without losing discriminative information and with less computational complexity. The goal of algorithms is not for cleaning or the noise removal of the corrupted speech signal but the feature itself is more robust to noise. No noise estimation is needed, and the algorithms utilize knowledge of the underlying speech spectrum. They designed two different processing methods such as harmonic demodulation and peak- to valley ratio
locking to minimize mismatch between clean and noisy speech features. They have tested these proposed methods against normal MFCC feature extraction techniques. They achieved good results for different SNRs using the proposed methods. The accuracy of speech recognition is about 60% for different SNRs in noisy environment using MFCC technique. For a database of TI46 isolated digits, the performance is improved from 60% to 95% and for Aurora2 connected digit string database performance improved from 58 to 80%.

Thuong Le-Tien and H. Dinh Chien (2004) presented a wavelet-neural network based method for Vietnamese Speech Recognition applied to Robot Communications [30]. The Vietnamese language is a difficult language for speech synthesis and recognition. Formants and pitch periods of speech signals were extracted using wavelet transforms as speech features and neural networks were used for matching recognition. The result of the practical modal of robot control was demonstrated with Vietnamese speech through some fundamental commands for the testing of the proposed approach. The interfacing between the robot and computer were programmed in Matlab 5.3 and Turbo C++ languages.

Jingdong Chen et al. (2004) investigated the spectral sub band centroid (SSC) features for speech recognition [31]. It is demonstrated that in clean speech conditions SSCs can produce performance comparable to that of MFCCs provided that the number of sub bands are properly selected. They have used an isolated spoken word database that was designed and collected by Texas Instruments (TI) i.e. TI46 database for this task, this database consists of speech samples from 16 speakers, including eight male and eight female speakers. The vocabulary consists of ten isolated digits from 0 (zero) to 9 (nine), and 26 isolated English alphabet letters from “a” to “z,” and ten isolated words (‘enter’, ‘erase’, ‘go’, ‘help’, ‘no’, ‘rubout’, ‘repeat’, ‘stop’, ‘start’, ‘yes’). There are 26 utterances of each word from each speaker among which ten of these are designated as training, and the remaining 16 are designated as testing tokens. The speech signal is digitized at a sampling rate of 12.5 kHz. The recognition system used an HMM-based multi speaker isolated speech recognizer. Eight states were used for each model. Mixtures of eight multivariate Gaussian distributions with diagonal covariance matrices were used for each state to approximate its probability density function. Speech samples were analyzed for every 10 ms with a frame
width of 32 ms, pre emphasized and Hamming windowed. They showed that the new dynamic SSC coefficients are more resilient to noise than the MFCC features.

Rafik Djemili et al. (2004) proposed an algorithm for Arabic isolated digit recognition [32]. The input vector consists of twelve Mel-frequency Cepstral coefficients and log of the energy used as speech features. The temporal features of the speech signal were extracted using Hidden markov models (HMM) and multi-layer perceptrons back-propagation algorithm was used to train the networks. The proposed algorithm was evaluated with limited database of isolated digits 0 to 9 in speaker independent manner. Training set consisting of twenty occurrences of each digit by 20 talkers (10 male and 10 female). Two different test categories conducted, firstly they used the same 20 talkers as were used in the training then the new set of 6 talkers with five occurrences per digit per talker were used giving 300 occurrences of digits (0-9). Their results showed that the word error rates significantly reduced comparing with an HMM word recognition system.

Teddy Surya Gunawan and Eliathamby Ambikairajah (2004) presented a fractional bark gammatone filter based on a short-term temporal masking threshold to noise ratio (MNR) for enhancement of noisy speech [33]. The noisy corrupted signal is first divided into number of sub bands using fractional bark accuracy, and then these sub-band signals weighted in the time domain according to a short-term temporal masking threshold to noise ratio estimate. The performance of the proposed algorithm was compared with other standard speech enhancement methods over six different noise types and three different SNRs. They showed that the proposed algorithm gives good results for speech enhancement applications for different environmental noises.

Chin Luh Tan and Adznan Jantan (2004) investigated feed-forward multi-layer perceptrons trained by back-propagation in spoken digit recognition [34]. They proposed an automated system from the training stage to the recognition stage without the need of manual cropping for speech. Speech features were extracted using Linear predictive coding (LPC) and multilayer perceptrons (MLP) with back-propagation used for training and designing the pattern classifiers. The analysis, design and development of the automation system is implemented in MATLAB. Accuracy of more than 95% was achieved for unknown pattern (spoken by unknown speakers). The results also showed
that the performance of the network was improved when more training datasets used to
train the network.

Panu Somervuo and Teuvo Kohonen (2004) constructed Self-Organizing Map (SOM) and Learning Vector Quantization (LVQ) algorithms for variable-length and warped feature sequences [35]. The Self-Organizing Map (SOM) consists of a two-dimensional grid in which the distribution of input items projected nonlinearly. The models associated with the grid points were taken as sequences of real feature vectors. Dynamic time warping (DTW) was used to compute the distances between feature sequences. The DTW-SOM was used for unsupervised clustering applications like hand writing recognition and DTW-LVQ was used in speaker-independent isolated-word recognition and achieved a test-set error rate of 1.5 %.

Florian Honig et al. (2005) developed revised processing steps for Perceptual Linear Prediction (PLP) that combines the advantages of both MFCC and PLP [38]. Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the most popular acoustic features extraction techniques used in speech recognition depends on the speech recognition task. The main differences between PLP and MFCC lie in the filter-banks, the equal-loudness pre-emphasis, and the intensity-to-loudness conversion and in the application of Linear Prediction (LP). The most prominent difference between them is the shape, the number and the width of the filters. The Bark filter-bank consists of 19 asymmetrically shaped filters while the Mel filter-bank contains typically 24–40 triangular filters which have a 50%-overlap. They have improvised the filter-bank, the equal-loudness pre-emphasis and the input for the linear prediction. They have shown that the new variant of PLP performs better than both MFCC and conventional PLP for a wide range of clean and noisy acoustic conditions especially for a broadcast news transcription task and a corpus of children’s speech.

Xiaodong Cui and Abeer Alwan (2005) proposed and implemented feature compensation based on Polynomial regression of signal to noise ratio (SNR) for noise robust automatic speech recognition (ASR) [39]. The bias difference between clean and noisy speech features is approximated by a set of polynomials which are estimated from adaptation data from the new environment by expectation-maximization (EM) algorithm under the maximum likelihood (ML) criterion. Firstly, the utterance SNR for the speech
signal is estimated and then noisy speech features are compensated by regression polynomials. Acoustic HMMs trained with clean data and compensated speech features are decoded. The experiments conducted for different noises and the results were compared on the Aurora 2 (English) and the German part of the Aurora 3 databases. They showed that the results using the Aurora 3 database gives the best performance for well-matched, medium-mismatched and high-mismatched conditions, respectively.

Manal El-Obaid et al. (2006) presented a paper on recognition of isolated Arabic speech phonemes using artificial neural networks [40]. Features of the Arabic speech phoneme were extracted using Cepstral coefficients with hamming window of frame size 512 samples and 170 overlapping samples. Supervised learning method of Multi Layer Perceptron Neural Network (MLP) based on Feed Forward back propagation was used for recognition of Arabic speech phonemes. 34 phonemes of Arabic language from KAPD (King Abdul Aziz Phonetics Database) were used for experiments and achieved a recognition rate around 96.3% for most of the phonemes.

Patricia Melin et al. (2006) described the use of neural networks in voice recognition [41]. Speaker recognition can be divided into speaker identification and speaker verification. Speaker recognition systems used to verify the speaker identity and control access to different devices like remote access to computers. They experimented different techniques such as Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms and achieved good results for the modular neural approach.

Engin avci (2006) demonstrated an automatic recognition system for Turkish words using discrete wavelet neural networks based on adaptive entropy [42]. They implemented Discrete Wavelet Neural Network (DWNN) model which consists of discrete wavelet layer used for adaptive feature extraction in the time-frequency domain and multi-layer perceptron feed-forward neural network layer used for classification. The speech recognition performance of this method was demonstrated for a total of 20 individual speakers (10 male and 10 female) and was evaluated using noisy Turkish word signals and the results showed the effectiveness of the automatic system. The rate of correct recognition achieved was about 92.58%.
Iosif Mporas et al. (2007) compared different feature extraction techniques for the task of speech recognition [43]. They proposed discrete Fourier transform (DFT) and discrete wavelet packet transforms (DWPT) based speech parameterization methods against traditional techniques such as the Mel-frequency Cepstral coefficients (MFCC) and perceptual linear predictive (PLP) Cepstral coefficients. They have demonstrated that the widely-used Mel frequency Cepstral coefficients are not the most appropriate choice of parameters when maximization of the absolute speech recognition performance is desired.

R. Schluter et al. (2007) introduced an acoustic feature set based on a Gammatone filter bank for large vocabulary speech recognition [44]. Combination of Gammatone features with number of standard acoustic features like MFCC (Mel frequency Cepstral coefficients), PLP (Perceptual Linear Prediction coefficients) and MF-PL-P were experimented and obtained good results. They achieved a relative improvement of about 12% in word error rate compared to the best single feature system.

Nima Yousefian and Morteza Analoui (2007) proposed a special model of radial basis probabilistic neural networks (RBPNN) as a classifier for speech recognition [45]. The proposed network has been tested on Persian one digit numbers dataset and produced significantly lower recognition error rate in comparison with other common pattern classifiers. Results showed that the MFCC features yield better performance compared to PLP features.

Khalid Saeed and Mohammad Kheir Nammous (2007) discussed a Speech-and-Speaker (SAS) Identification System for spoken Arabic digit recognition [46]. The standard format is used in this work, pulse-code modulation with a frequency of 22050 Hz, 16-bit mono. The speech signals of the Arabic digits from zero to ten were processed graphically using the algorithm of Toeplitz matrix minimal eigen values for signal-image description and feature extraction and Burg's estimation model used as classifying technique. All the experiments have proven that the image-based methods in speech recognition have given good recognition results when compared to other standard methods. They achieved the overall success rate was about 97.45% in recognizing one uttered word and identifying its speaker, and 92.5% in recognizing a three-digit password (three individual words).
Ji Ming et al. (2007) investigated the speaker identification and verification problem when speech signals are corrupted with environmental noises where the characteristics of the noise are not known [47]. The concept of speaker identification or verification utilized in biometric and other security devices which are generally mobile in nature, as the devices are movable and the effect of noise sources are time-varying and unpredictable. They mainly focused in implementing the model for real time applications. They have implemented three different models based on sub band feature formats, BSLN-Cln system trained on clean data, BSLN-Mul system trained on simulated multi condition data and the proposed system which is combination of multi condition model training with missing-feature theory. They have used two different databases to test their proposed algorithm. Firstly TIMIT database was chosen and rerecorded the speech data in the presence of various noise types and used the same to test the model for speaker identification with a focus on the varieties of noise, then hand held-device speech data was collected in realistic noisy conditions and validated the model for real-world speaker verification. The proposed model is compared to baseline systems and achieved lower error rates. They showed that the combination of multi condition training with a missing-feature model reduces the mismatch between training and testing speech data improved the robustness of the model.

Ahsanul kabir and Sheikh mohammad masudul ahsan (2007) proposed Vector Quantization in Text dependent automatic speaker recognition using Mel-frequency Cepstrum coefficients [48]. Speaker recognition or identification technique is used to verify the speaker’s voice identity and control access to the different security services such as biometric security system, voice dialing, telephone banking, telephone shopping, database access services, information services, voice mail, and security control for confidential information areas and remote access to the computers. Firstly they created a codebook of the speakers to characterize their vocal characteristics using training sentences and then they compared a sample of a speaker's voice to the codebook to determine the identity of the speaker. Different windowing techniques such as Triangular, Hanning and Hamming windows were used to test the recognition rate in linear scale of the system by varying the number of centroids up to sixteen. They achieved good results with Hamming window and the number of centroids is chosen for more than eight. They have taken 32 samples of different speakers and the system algorithms are programmed in MATLAB environment.
Yang Shao and De Liang Wang (2008, 2009) proposed novel based auditory features for robust speaker identification [50, 53]. Human listeners are able to segregate and recognize speech in noisy conditions but the automatic Speaker recognition systems performance drops in noisy environment. Any speech or speaker recognition system mainly consists of three stages: feature extraction of speech signals, speech or speaker modeling, and decision making using pattern classification methods. They have explored different feature dimensions and incorporated dynamic features towards improving the performance of robust speaker recognition system using a front-end based on computational auditory scene analysis. They found that one of the auditory features known as gamma tone frequency Cepstral coefficients performed substantially better than a conventional speaker features. Gammatone features are derived by applying a Cepstral analysis on Gammatone filter bank responses. Their designed recognition system achieved significant performance improvements compared with an advanced front-end in a wide range of signal-to-noise conditions.

Mohamad Adnan Al-Alaoui et al. (2008) have compared two different classifier methods for automatic Arabic speech recognition for isolated words and sentences [51]. Linear predictive coding (LPC) is used for extracting Cepstral features from the speech signals, two different classifier techniques such as standard Hidden Markov Models (HMM) and Al-Alaoui Algorithm based Neural Networks used for pattern training and classification. Multilayer perceptron (MLP) with two layers of neurons, and back propagation method using Al-Alaoui algorithm is used for modeling the proposed neural networks. The database used to train the models and testing contains Arabic language teaching lessons available in the books for the illiterates and collection of uttered words supplied by the Lebanese Ministry of Education. The proposed method gave significant results compared to the already implemented HMM method for the recognition of words, and it has overcome HMM in the recognition of sentences.

Yang Xiaocui and SUN Lihua (2009) implemented HMM based English Speech Recognition System in MATLAB. They simulated around 100 common English sentences using MATLAB Voice tool box [52]. They have chosen conventional MFCC coefficients as speech features extraction technique. They have collected speech samples from five male speakers and five female speakers for training the system. This speech recognition system is utilized in Tourism English training, and the pronunciation accuracy is tested as a result of speech recognition displays in scores.
Meysam Mohamad pour and Fardad Farokhi (2009) presented an advanced method for Persian language speech recognition to classify speech signals with the high accuracy at the minimum recognition time [55]. In this method the recorded signal was preprocessed which includes denoising with Mels Frequency Cepstral analysis and feature extraction using discrete wavelet transform (DWT) coefficients, then these features are applied to Multilayer Perceptron (MLP) network for training and classification. Dynamic time warping (DTW) was used to find whether two speech waveforms represent the same spoken phrase. Different parameters like speaker dependency, discrete or continues spoken word/sentences, size of vocabulary, language constrains, noisy environment, incompatibility between train and test conditions, dissimilarity in expressing and pronouncing same word by different speakers and one speaker pronouncing the same word differently at different times affect the accuracy of speech recognition systems. They presented and developed an automatic Persian speech recognition system using UTA algorithm with increased system learning time from 18000 to 6500 epochs and achieved average system accuracy to 98% approximately.

M.A.Anusuya and S.K.Katti (2009) provided a comprehensive survey of research work on speech recognition [56]. They mentioned significant progress has been made in the last two decades and different areas of speech recognition research work to be focused on. They opined that a robust speech recognition system should be effective under full variation in environmental conditions, speaker variability etc.

Okko Johannes Räsanen et al. (2009) developed a computationally clustering algorithm for vector quantization (VQ) of speech signals for the task of unsupervised pattern discovery (PD) from speech signals [57]. In this paper, they have discussed convergence of the clustering algorithm and the ultimate quality of VQ codebooks. The method is especially designed for incremental learning problems where the size of the codebook and the amount of input material is difficult to determine beforehand but when some resolution limits are known. In the clustering process, new clusters are created if they are not sufficiently similar to existing ones. It was observed that there was no much improvement in the proposed SLVQ algorithm when compared to more sophisticated k-means algorithm. Mainly the overall size of the codebook has a significant impact on the quality of the learned patterns and therefore on the recognition accuracy.
Revathi A et al (2009) presented a paper on Text independent Speaker recognition and Speaker independent speech recognition using iterative clustering approach [58]. They have experimentally evaluated clustering approach method on the set of isolated digits and continuous speeches from TI digits_1 and TI_digits_2 database for speech recognition and on speeches of 50 speakers' randomly chosen from TIMIT database for speaker recognition. They proposed robust perceptual features and iterating clustering approach in clean speech environment.

Dr. Mustapha Guezouri et al. (2009) have been motivated to investigate neural networks as a speech recognition tool [59]. Temporal Radial Basis Functions (TBRF) was analyzed in the speech recognition oriented vowel classifications. They reported good recognition success rates in identifying the vowels taken from natural speech samples from the TIMIT corpus of American speech. They have demonstrated with different examples and showed that the mentioned neural network could able to learn acoustic-phonetic features, such as formant movements and segmentation, and use them effectively as internal abstractions of speech.

Lindasalwa Muda et al. (2010) implemented voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques [61]. They divided the Voice Recognition Algorithms into two phases, Training phase and testing phase. They described the different steps in extracting the MFCC feature extraction process, the different steps includes Pre-emphasis, Framing, Hamming Windowing, Fast Fourier transform, Mel filter bank processing, Discrete Cosine Transform, and delta energy and delta spectrum. Non linear Dynamic Time warping technique was used for features matching. Gold wave is used for recording the test speech samples.

Jayanta Kumar Basu et al. (2010) submitted their review on use of Artificial Neural Network in Pattern Recognition [62]. They summarize and compare some of the well-known methods used in various stages of a pattern recognition system using Artificial Neural Networks (ANNs). They opined that the design of any recognition system requires careful on definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and
learning, selection of training and test samples, and performance evaluation. The emerging applications like face recognition, and cursive handwriting recognition, require robust and efficient pattern recognition techniques.

Md. Rabiul Islam and Md. Fayzur Rahman (2010) presented a paper on Noise Robust Speaker Identification using PCA based Genetic Algorithm [64]. This paper emphasized speaker identification system on Principal Component Analysis (PCA) based Genetic Algorithm that deals with detecting a particular speaker from a known population under noisy environment. They have eliminated the noise from speech utterance using wiener filtering technique. Different types of feature extraction techniques such as Real Cepstral Coefficients (RCC), Mel Frequency Cepstral Coefficients (MFCC), Delta Mel Frequency Cepstral Coefficients (ΔMFCC), Delta Delta Mel Frequency Cepstral Coefficients (ΔΔMFCC), Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC) have been experimented to improve the performance of the text dependent speaker identification system under noisy environment. ΔMFCC feature extraction technique achieved the highest identification success rate 85.73%. The system has some limitations while testing with NOIZEOUS speech database, vocabulary was limited and the numbers of users were limited.

Wouter Gevaert et al. (2010) investigated speech recognition classification performance using two standard neural networks such as Feed-forward Neural Network (NN) with back propagation algorithm and Radial Basis Functions Neural Networks [65]. Though there was a good progress in speech recognition since last 50 years, the field is still facing lot of problems. Some of the common issues that impacts the performance of speech recognition systems are Speaker variation (The same word is pronounced differently by different people because of age, sex, anatomic variations, speed of speech, emotional condition of the speaker and dialect variations), Background noise (environment noise or the noise added by speaker himself), Influence of intonation and putting stress on syllables and other external factors like position and direction of the microphone in respect to the speaker. They observed poor classification when pre-processing stage processed with spectrogram combined with entropy based endpoint detection, then they experimented Mel Frequency Cepstrum Coefficients feature extraction technique for the pre-processing stage and achieved good classification results.
Both the Multilayer Feed forward Network with back propagation algorithm and the Radial Basis Functions Neural Network achieved satisfying results with Mel Frequency Cepstrum Coefficients. They observed that the pre-processing quality gave the biggest impact on the neural networks performance.

Suma S.A and Dr. K.S. Gurumurthy (2010) described a computationally simple Pitch extraction algorithm using Average Magnitude Difference Function (AMDF) using weighted autocorrelation which is very useful for accurate Pitch Period extraction for LPC Speech Coders in Noisy Environments [66]. They analyzed the speech signals with gammatone filter bank that splits the full band speech signal into sub bands and pitch was extracted for each sub band of speech signal, then Signal to Noise Ratio (SNR) was determined for each sub band and then the average of pitch periods of the highest SNR sub bands is calculated and obtained optimal pitch value. They compared the effectiveness of the proposed AMDF using weighted Autocorrelation and the existing method and concluded that the new approach AMDF using weighted autocorrelation performed well in different environmental noises.

Dr. E.G. Rajan and Santyanand Singh (2010) proposed Vector Quantization approach for text independent speaker recognition using MFCC and inverted MFCC features [67]. They used Gaussian shaped filter (GF) for calculating MFCC and inverted MFCC in place of traditional triangular shaped bins combined with Vector Quantization (VQ) feature matching technique. The performance of both MFCC and inverted MFCC improved with GF over traditional triangular filter (TF) based implementation. They achieved 98.57% of efficiency with a very short test voice sample of 2 seconds with the proposed method.

Fu Guojiang (2011) proposed a Novel Isolated Speech Recognition based on Neural Networks [68]. Recognition of the words was carried out in speaker dependent mode and has used same data for both training and testing purpose. He has chosen 16 Linear Predictive Cepstral coefficients with 16 parameters from each frame as feature extraction. He compared the proposed Radial Basis Function Neural Network with MLP (multi-layer perceptron) classifier and shown RBF was more suitable for the recognition of isolated words.
Fatma zohra Chelali et al. (2011) have investigated Speaker Identification System based on PLP Coefficients and Artificial Neural Networks [69]. They have chosen Perceptual Linear Prediction (PLP) coefficients as feature extraction technique. PLP feature extraction technique majorly consists of three phases, Critical band spectrum selection, equal loudness curve, and intensity power law. Multilayer perceptron (MLP) artificial neural networks have been chosen for training the models. They designed a speaker identification and phoneme classification system and experimented using 14 Arabic phonemes, specifically the Arabic fricatives uttered by 4 Algerian native speakers and achieved good recognition rate using PLP-MLP algorithm.

Xing Fan and John H. L. Hansen (2011) developed a seamless speaker identification system for whispered speech audio streams [71]. The performance accuracy of Speaker identification/recognition systems trained with neutral speech degrades significantly when the identification/recognition is performed for whispered speech as there is a profound difference between whispered and neutral speech in both excitation and vocal tract function. They designed a closed-set speaker ID system based on an Mel-frequency Cepstral coefficient-Gaussian mixture model (MFCC-GMM) and observed performance degradation for whisper speaker, then an alternative feature extraction algorithm based on linear and exponential frequency scales were applied and two-dimensional feature space was proposed to predict whispered utterances. With this they achieved absolute improvement in speaker recognition around 8.85% to 10.30% and obtained speaker ID performance around 88.35% for a total of 961 read whisper test utterances.

Dimitrios Dimitriadis and Petros Maragos (2011) examined the contribution of energy computation and filter bank design to the overall front end robustness when the investigated features are applied to noisy speech signals, in mismatched training-testing conditions [72]. They showed in their earlier work that Teager energy Cepstrum coefficients (TECC) s are more robust to noise and exhibit improved performance compared to the widely used Mel frequency Cepstral coefficients (MFCCs). In the present work, they investigated the connection between the filter bank design (the filter shape and bandwidth), the energy estimation scheme and the automatic speech recognition (ASR) performance under a variety of additive and/or convolutional noise conditions.
Experimental results showed that selecting the appropriate filter bank and energy computation scheme would lead to significant error rate reduction over both MFCC and perceptual linear prediction (PLP) features for a variety of speech recognition tasks. A relative error rate reduction of up to 30% for MFCCs and 39% for PLPs was shown for the Aurora-3 Spanish task.

Cemal Hanilci and Figen Ertas (2011) investigated the impact of LP-residual cepstrum coefficients (LPRC) on speaker verification along with MFCC and linear predictive cepstrum coefficients (LPCC) [74]. They have chosen NIST 2001 database for speaker recognition evaluation (SRE). NIST 2001 SRE database contains 174 speakers (74 males and 100 females) with total number of 22,418 trials (2,038 genuine and 20,380 impostor trials). They have used Gaussian mixture model with universal background model (GMM-UBM) which was introduced by Reynolds, Quatieri, and Dunn in the year 2000. From the experiments, they showed that the LPRC features also useful in speaker verification as like MFCC and LPCC features and achieved good results with the combination of LPRC, LPCC, and MFCC features in pairs.

Mondher Frikha and Ahmed Ben Hamida (2012) compared the performance of ANN, Hybrid HMM and ANN Architectures for Robust Speech Recognition [75]. Two different kinds of Neural Networks (NN) such as Multilayer Perceptron (MLP) and Elman Recurrent Neural Networks (RNN) were investigated along with the hybrid connectionist-HMM systems. They tested the performance of the conceived systems using the TIMIT database in clean and noisy environments with two perceptually motivated features such as MFCC and PLP. They developed a preprocessing denoising stage based on wavelet transform and the robustness of the systems was evaluated with various types of additive noise at different SNR values.

Djellali Hayet and Laskri Mohamed Tayeb (2012) described different approaches for vector quantization in Automatic Speaker Verification [76]. They designed a novel architecture based on multiples codebook representing the speakers and the impostor model called universal background model (UBM) and compared the performance with another vector quantization approach used. They compared the proposed scheme with the baseline system, Gaussian Mixtures Models and Maximum Posteriori Adaptation. They achieved good results with Vector Quantization Universal Background Modal (VQ-
UBM) for 128 codebook size and improved the performance of vector quantization applied in speaker verification compared to baseline vector quantization. They observed that the size of codebook would influence the verification accuracy. They opined that the size of speech data should be increased in order to validate the experiments for larger databases.

Alfredo Maesa et al. (2012) implemented Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models [77]. They have extracted two different speech samples from 450 speakers which were randomly picked from the Voxforge.org audio database, one file is used to build the profile database and the second file is used to test the system performance. ASR is designed ASR system based on MFCC and GMMs. They achieved around 96% accuracy by the proposed approach. Algorithms were implemented in Matlab language and the time taken to recognize the speaker is about 2 seconds on a common PC. The ASR system is useful for real time security access control applications.

Reza Haghmaram et al. (2012) studied the automatic noise recognition (ANR) problem based on RBF and MLP neural networks classifiers using linear predictive and Mel-frequency Cepstral coefficients (LPC and MFCC) [78]. The proposed networks were evaluated on different types of stationary and non stationary noises with frame length of 20ms. The experimented results of networks were analyzed and compared to each other. They showed that the proposed ANRs of MLP network with LPC and RBF network with MFCCs achieved acceptable accuracy rate higher than 90%.

Addou DjaMel et al. (2012) introduced an efficient front-end for distributed Speech Recognition over Mobile [79]. A new set of feature vectors were proposed which have been estimated through three different steps namely extraction of speech signals from a de-noised acoustic frame using wiener filter and combining the Mel-Line Spectral Frequencies (MLSFs) coefficients with conventional MFCCs, optimizing the stream weights of multi-stream HMMs and then applying these features to Karhunen-Loeve Transform (KLT). They showed that the proposed system is compatible with 3GPP and 3GPP2 standards for both European (GSM) mobile and North American (CDMA) systems respectively. Recognition experiments were conducted on Aurora 2 connected digits database and results showed that the proposed front-end leads to a significant improvement in speech recognition accuracy for highly noisy GSM.
Lajish V. L et al. (2012) introduced a Nonlinear and artificial neural network speech model for speaker identification [80]. Reconstructed Phase Space (RPS) of the speech signal and the Phase Space Point Distribution (PSPD) parameters were used in modeling the speaker identity. Feed Forward Multi Layer Perceptron (FFMLP) classifier was used in their work. The PSPD features were obtained from five different vowels. The performance is compared by conducting the experiments repeatedly by taking different combination of PSPD, MFCC, pitch and first formant frequency. The experimental results indicated that the proposed phase space approach by itself is still below (31.60%) that of MFCC features (46.21%), but the combined approach in which the PSPD features, when used with MFCC, pitch and first formant frequency, offers enormous improvement in speaker identification (on an average of 83.40%) accuracy.

Hamdy K. Elminir et al. (2012) experimented different feature extraction techniques and analyzed the speech recognition evaluation parameters such as recognition success rate (%), training time, feature extraction time and PCA conversion time [82]. They evaluated different feature extraction techniques LPC, MFCC and ZCPA for continuous speech and compared between these feature extraction techniques to find the most suitable technique for speech recognition process, and tried to enhance the result by using PCA.

Mahmoud I. Abdalla et al. (2012) presented a paper on DWT and MFCCs based feature extraction method for Isolated Word Recognition [83]. A new set of features were generated by concatenating both DWT and MFCCs features. The MFCCs of the wavelet channels were calculated for capturing the characteristics of the speech signals and Neural Network used for training and classification. The results showed an improvement of the recognition rate of this new method over using MFCCs.

In the research work undertaken, studies were carried out for better performance of ASRs in noisy environment for different combinations of speech feature extraction and modeling techniques which are explained in the subsequent chapters.