CHAPTER 3

LITERATURE SURVEY

In this chapter different speaker and speech recognition techniques described in the literature are reviewed. The literature is listed year wise under different techniques proposed.

3.1 INTRODUCTION

Spoken language is an important means of human communication. Speaker recognition is a useful tool for authentication. Human computer interaction requires machine to recognise and understand speech to provide a comfortable and natural communication.

Speech recognition software is used in a wide range of applications, from automated commercial phone systems to enhancing personal productivity. It helps the computer to recognize general and naturally flowing utterances from a wide variety of users. This technology appeals to anyone who needs hands-free approach to complete tasks. The goal is to guarantee user-friendly interface for any task.

Joseph (1996) describes that the modern speech understanding systems merge interdisciplinary technologies viz., Signal Processing, Pattern Recognition, Natural Language, and Linguistics into a unified statistical framework. Digital signal processing, vector-oriented and linear algebra-based processors dominate this
area of research. The current generation of DSP-based systems rely on sophisticated statistical models implemented using a complex software paradigm. Such systems are now capable of understanding speech input for large vocabulary in operational environments.

3.2 EVOLUTION OF SPEECH RECOGNITION

Melanie (2011) describes the history of speech recognition as follows.

- **1950s and 1960s: Baby Talk**
  In the beginning, speech recognition systems could understand only digits. In 1952 Bell laboratories designed “Audrey” system, which recognized digits spoken by a single voice. In 1962 IBM demonstrated at the World’s Fair its “Shoebox” machine, which could understand 16 words spoken in English.

  Labs in the United States, Japan, England and the Soviet Union developed other hardware dedicated to recognizing spoken sounds, expanding speech recognition technology to support four vowels and nine consonants. They may not sound much today but these efforts were an impressive start especially when you consider how primitive computers themselves were at that time.

- **1970s: Speech Recognition Takes Off**
  Speech recognition technology made major strides in 1970s. From 1971 to 1976, DARPA’s Speech Understanding Research (SUR) program was one of the largest in speech recognition under the Department of Defence (DoD). It led to development of Carnegie Mellon’s “Harpy” speech-understanding system. Harpy could understand 1011 words. It was significant because it introduced a more efficient search approach called “beam search” to prove the finite-state network of possible sentences.
The 1970s also marked a few other important milestones in speech recognition technology including founding of the commercial speech recognition companies first of its kind - Threshold technology and Bell Laboratories introducing a system that could understand multiple speakers.

- **1980s: Speech Recognition Turns toward Prediction**
Over the next decade, recognition vocabulary size increased from about a few hundred words into several thousand words by using a new statistical method known as the hidden Markov model. It had the potential to recognize unlimited number of words.

However, whether speech recognition software at the time could recognize 1000 words or 5000-word vocabulary as IBM’s system did, a significant hurdle remained. These programs worked only with discrete words as one had to pause after each and every word.

- **1990s: Automatic Speech Recognition comes to the Masses**
In the 1990s computers with faster processors finally arrived and speech recognition software became commercially viable.

In 1990, Dragon launched the first consumer speech recognition product, Dragon Dictate. Seven years later, the much-improved Dragon Naturally Speaking arrived. The application recognized continuous speech.

In 1996, BellSouth introduced a dial-in interactive voice recognition system. It was supposed to give you information based on what you said on the phone.


- **2000s: Google comes along**

By 2001, computer speech recognition had topped out at 80% accuracy and near the end of the decade; the technology’s progress seemed to be stalled. Speech recognition technology development began to edge back into the forefront with one major event: the arrival of the “Google Voice Search app for the iPhone”. In 2010, Google added “personalized recognition” to Voice Search on Android phones, so that the software could record users’ voice searches and produce a more accurate speech model. The company also added Voice Search to its Chrome Browser in mid-2011. Relying on cloud-based processing Apple introduced “Siri”. It infers from what it knows about you and generates a contextual reply based on your voice input.

- **The Future: Accurate, Ubiquitous Speech**

The explosion of voice recognition applications indicates that speech recognition’s time has come and that you can expect plenty more applications in the future. These applications will not only let you control your PC by voice or convert voice to text, they will also support multiple languages, offer assorted speaker voices for you to choose from and integrate into every part of your mobile devices.

Having considered the evolution of speech recognition system, we now describe the development of speech recognition system by different researchers in detail.

### 3.3 SPEECH RECOGNITION SYSTEM

Rabiner & Juang (2006) describes a block diagram of a speech recognizer (figure 3.1) that follows the Bayesian framework. The recognizer consists of three
processing steps: feature analysis; pattern matching, and confidence scoring. It also consists of three trained databases: the set of acoustic models, the word lexicon, and the language model.

Figure 3.1 Speech Recognizer (adapted from: Rabiner & Juang 2006)

Three required steps were described to define, train, and build an ASR system as given below.

- Step 1: Choose the recognition task.
- Step 2: Train the models.
- Step 3: Evaluate recognizer performance.

3.4 SPEECH RECOGNITION TECHNIQUES

Santosh et al (2010) say that the goal of speech recognition is the ability of the machine to "hear," understand," and act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell
Laboratories, where an isolated digit recognition system for a single speaker was developed. The goal of automatic speaker recognition is to analyse, extract characteristics and recognize information about the speaker identity. There are four stages in speaker recognition system.

1. Analysis
2. Feature extraction
3. Modeling
4. Testing

### 3.4.1 Speech Analysis Techniques

The speech analysis technique is done using any of the three techniques specified in the table 3.1 as per Santosh et al (2010).

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Method</th>
<th>Property</th>
<th>Algorithm used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Segmentation Analysis</td>
<td>Extracts speaker information by processing in blocks using the frame size and shift in the range of 10-30 ms</td>
<td>LPCC or MFCC with GMM</td>
</tr>
<tr>
<td>2</td>
<td>Sub Segmental Analysis</td>
<td>Extracts speaker information by processing in blocks using the frame size and shift in the range of 3-5 ms</td>
<td>LPCC or MFCC with GMM</td>
</tr>
<tr>
<td>3</td>
<td>Supra Segmental Analysis</td>
<td>To analyze the characteristic behavior of the speaker by processing in blocks using the frame size of 250ms</td>
<td>LPCC or MFCC with GMM or using the concept of instantaneous pitch</td>
</tr>
</tbody>
</table>
State of the art speaker recognition systems mostly use vocal tract related speaker information represented by the spectral or cepstral features like linear prediction cepstral coefficients (LPCC) or mel frequency cepstral coefficients (MFCC). These features provide good recognition performance. They nearly represent complete vocal tract information i.e. LPCC or MFCC captures the formants and their bandwidth information characterizing the vocal tract completely, but pitch is only one aspect of speaker information due to source.

LP residual is a result of passing the speech signal through inverse filter (i.e., removing the vocal tract information). This is approximately equal to the excitation signal or source information.

Vocal-tract features are calculated by using Linear Predictive coding. So these are the conventional features, which are added with the above features (sub segmental, segmental & supra segmental). Hence in the speaker recognition systems the extraction of complete source information can be improved by the time domain analysis of the LP residual.

The excitation source information is extracted by processing the LP residual in the time domain. The time domain processing of the LP residual is computationally intensive. LP residual can be processed from the other domains like frequency or cepstrum. This has to be done by keeping in view the blocking effect that is present in these domains (Ravi et al 2011).

Ladan & Douglas (2009) investigated a non-parametric classification of English phonemes in speaker-independent continuous speech. The system employs a powerful and intuitive non-parametric classifier. The recognition result shows a
promising increase in the percentage of correctness over the conventional HMM-based phoneme recognition. In addition, applying the approximate nearest neighbour approach for the classification purpose rather than the exact one, leads to achieving a very lower training execution time compared to the HMM-based system, and also a comparable execution time for the testing. The outcome was a considerable reduction in the $k$-NN search space and hence the execution time, and also a slight increase in the recognition performance.

3.4.2 Feature Extraction Techniques

The speech feature extraction is used to reduce the dimensionality of the input vector while maintaining the discriminating power of the signal from fundamental formation of speaker identification and verification system. The number of training and test vector needed for the classification problem grows with the dimension of the given input. Hence we need feature extraction of speech signal (Santosh et al 2010).

Following are some feature extraction techniques for data classification and dimensionality reduction.

1. Principal Component analysis (PCA): PCA is a powerful technique for extracting structure from possibly high-dimensional data sets (Tetsuya & Yasuo 2007).

   Stiphane & Leila (2000) presents a fast speaker adaptation technique dedicated to automatic speech recognition systems using artificial neural networks (ANNs) for hidden Markov models (HMMs) state probability estimation. With only 20 words of adaptation data, results show a 25% relative decrease of the word error rate over the speaker independent system, and a 15% decrease over the standard affine transformation adaptation approach.
2. **Linear Discriminant Analysis (LDA):** The use of Linear Discriminant Analysis for data classification is applied to classification problem in speech recognition. The prime difference between LDA and PCA is that, the PCA does more of feature classification and LDA does data classification (available from: Balakrishnama & Ganapathiraju).

3. **Independent component analysis (ICA):** In this method the goal is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation (Hyvärinen & Oja 2000).

4. **Linear predictive coding (LPC):** It is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters (available from: Linear Predictive Coding).

One of the more powerful analysis techniques is the method of linear prediction. Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech.
In reality the actual predictor coefficients are never used in recognition, since they typically show high variance. The predictor coefficients are transformed to a more robust set of parameters known as cepstral coefficients (available from: Feature Extraction).

Kuah et al (1994) conducted a text-independent voice recognition experiment using an artificial neural network. The speech data were collected from three different speakers uttering thirteen different words. Each word was repeated ten times. The speech data were then pre-processed for signal conditioning. A total of 12 feature parameters were obtained from Cepstral coefficients via a Linear Predictive Coding (LPC). These feature parameters then served as inputs to the neural network for speaker classification. A standard two-layer feed forward neural network was trained to identify different feature sets associated with the corresponding speakers. The network was tested for the remaining unseen words in text independent mode. The results were very promising with a voice recognition accuracy of more than 90%.

Chen et al (2008) presented an effective method for speaker identification system. Based on the wavelet transform, the input speech signal is decomposed into several frequency bands, and then the linear predictive cepstral coefficients (LPCC) of each band are calculated. In this study, the effective and robust LPCC features were used as the front end of a speaker identification system. In order to effectively utilize these multi band speech features, a multi-band 2-stage VQ was proposed as the recognition model. Different 2-stage VQ classifiers were applied independently to each band, and then errors of all 2-stage VQ classifiers were combined to yield total error. The experimental results show that the proposed method is more effective and robust than the baseline models proposed previously.
5. **Cepstral Analysis:** Speech is composed of excitation source and vocal tract system components. In order to analyze and model the excitation and system components of the speech independently and also use that in various speech processing applications, these two components have to be separated from the speech. The objective of cepstral analysis is to separate the speech into its source and system components without any a priori knowledge about source and / or system (available from: Cepstral Analysis of Speech).

Prasad et al (2001) used Artificial Neural Networks as research tool to accomplish Automated Speech Recognition of normal speech. A small size vocabulary containing the words YES and NO is chosen. Spectral features using cepstral analysis are extracted per frame and imported to a feed forward neural network, which uses a back propagation with momentum training algorithm. The network is trained to recognize and classify the incoming words into the respective categories. The output from the neural network is loaded into a pattern search function, which matches the input sequence with a set of target word patterns. The level of variability in input speech patterns limits the vocabulary and affects the reliability of the network. The results from the first stage of this work are satisfactory and thus the application of artificial neural networks in conjunction with cepstral analysis in isolated word recognition is promising. The system provided satisfactory results. It is robust enough to account for a speaker independent input. Though the encouraging success of the current system is achieved based on a limited vocabulary, the system can be expanded to a larger vocabulary by extending the number of subnets used in the architecture. The key solution is to increase the number of features extracted on each frame at the cost of additional processing time.
6. **Mel frequency cepstral coefficient:** It is based on signal decomposition with the help of a filter bank, which uses the Mel scale expressed on the Mel-frequency scale. The MFCC is the result of a discrete cosine transform of the real logarithm of the short-term energy. Mel scale cepstral analysis is very similar to perceptual linear predictive analysis of speech, where the short-term spectrum is modified based on psychophysically based spectral transformations. In this method, the spectrum is warped according to the MEL scale, where as in PLP the spectrum is warped according to the Bark scale. The main difference between Mel scale cepstral analysis and perceptual linear prediction is related to the output cepstral coefficients. The output cepstral coefficients are then computed based on this model. In contrast Mel scale cepstral analysis uses cepstral smoothing to smooth the modified power spectrum. This is done by direct transformation of the log power spectrum to the cepstral domain using an inverse Discrete Fourier Transform (IDFT). The MFCC has good performances in speech recognition (available from: Feature Extraction).

Aranda et al (2005) proposed an environmental sounds recognition system using LPC- Cepstral coefficients for characterization and a back propagation artificial neural network as verification method. The verification percentage was 96.66% although the number of feature vectors was small; specifically two feature vectors were used. The lowest percentages were obtained for noisy sound sources, as car, motorcycles and airplanes.

Zhongming (2010) proposed a key word detection method for continuous speech in noisy environment. In the proposed method, the widely used energy, zero crossing, entropy and MFCCs were extracted to generate an audio feature set. Robust endpoint detection algorithm is also used which makes the feature modify its parameter by adapting to the strength of background noise. Then HMMs are used for the classifiers. Experiments were made under different types of noises and the
results show that this method is more accurate and more anti-noise than traditional methods. This method was used in a student management system to recognize some key words.

Soon (2011) analyzed the voice recognition algorithm based on HMM (Hidden Markov Model) in detail. The feature vector of each voice characteristic parameter is chosen by means of MFCC (Mel Frequency Cepstral Coefficients). The extracting algorithm of syllable parts from continuous voice signal is introduced. It shows the relationship between recognition rates and number of applying syllables and number of groups for applying syllables. The core engine of the HMM method is described, and simple syllables were used for the recognition process. In order to achieve a high recognition rate for different syllables, significant quantitative information of syllables is required. MFCC parameters were used. MFCC with a mel frequency index of 24 provides a higher recognition rate (96% per 72 syllables). Speaker dependent recognition requires only a mel frequency index of 14 during training in comparison to the 24 required for speaker independent recognition training. Based on the results of this study, more words can be added frequently to the database. By increasing the number of voice samples being trained, HMM can be widely applied to real life applications and, ultimately, a voice recognition system can be produced.

7. **Filterbank Analysis:** The human ear resolves frequencies, non-linearly across the audio spectrum and, empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. A popular alternative to linear prediction based analysis is therefore filterbank analysis since this provides a much more straightforward route to obtain the desired non-linear frequency resolution. However, filterbank amplitudes are highly correlated
and hence, the use of a cepstral transformation in this case is virtually mandatory, if
the data is to be used in a HMM based recogniser with diagonal covariances
(available from: Filterbank Analysis).

Dimitrios et al (2011) discussed about how energy computation and filter
bank design contribute to the overall front-end robustness, especially when the
investigated features are applied to noisy speech signals, in mismatched training-
testing conditions. Theoretical and experimental results showed that the filter
bandwidth and the shape of the spectrum are the most important factors affecting
speech recognition performance in noise. For large filter bandwidths, the Teager–
Kaiser operator outperforms (on the average and for most noise types) than the
squared amplitude energy computation scheme for speech recognition in noisy
conditions.

Experimental results show that selecting the appropriate filter bank and
energy computation scheme can 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 is shown for the Aurora-3 Spanish Task. The equivalent rectangular filter
bandwidths and the energy estimation scheme appear to be two of the most
significant parameters determining ASR performance.

ASR performance can be predicted for a particular choice of filter
bandwidth range and energy estimation scheme when the relative spectral energy
distributions of signal and noise are considered.

The generalized cepstral features are directly related to these energy
distributions. It is important to ensure a robust and efficient energy computation
process. The noisy cepstral coefficient deviations (deviations from the clean case) are, on average (RMS values), smaller than those of the MFCCs. This is due to the energy scheme and the wider filters employed.

The features using filters of different spectral shape present similar performance when their effective filter bandwidths are kept equal, regardless of their design parameters, for low and medium mismatch training/testing scenarios. For high mismatch, the energy computation scheme is usually the most important factor affecting performance. When advanced signal denoising and feature equalization techniques are applied in combination with the feature extraction scheme, the performance improvements appear to be additive on top of the signal and feature enhancement techniques, such as Wiener filtering and parameter equalization (PEQ). This is particularly important in building robust ASR systems.

8. **Spectral subtraction:** The background noise is the most common factor degrading the quality and intelligibility of speech in recordings. The noise reduction module intends to lower the noise level without affecting the speech signal quality. This module is based on the spectral subtraction performed independently in the frequency bands corresponding to the auditory critical bands.

The spectral subtraction method is a simple and effective method of noise reduction. In this method, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved. It is assumed that the signal is distorted by a wide-band stationary additive noise. The noise estimate is the same during the analysis and restoration, the phase is the same in the original and restored signal (available from: Noise Reduction).

9. **Cepstral Mean Subtraction (CMS):** It is one normalization method, which is to eliminate channel distortion in speech. It is based on the fact that any
convolutional distortion in the time domain transforms to additive distortion in cepstral domain. Though very simple, it shows effectiveness in applications. However, limited to its consumption that channel can be characterized as linear and time-invariant, CMS loses its effectiveness when the real environment cannot be adequately modeled like this.

Some modifications thus have been made to the conventional CMS. Two kinds of methods are popularly used. 1. To compensate channel distortion as nonlinear based on energy information at time domain, which is known as the two level CMS. 2. To renew the log spectrum of cepstral mean to be shaped more like a channel response and take the spectrum of transmission channel into account. MMCMNFWM is one of these instances (Pu et al 2004).

10. **Relative spectra filtering (RASTA):** To compensate for linear channel distortions the analysis library provides the ability to perform RASTA filtering. The RASTA filter can be used either in the log spectral or cepstral domains. In effect the RASTA filter band passes each feature coefficient. Linear channel distortions appear as an additive constant in both the log spectral and the cepstral domains. The high-pass portion of the equivalent band pass filter alleviates the effect of convolutional noise introduced in the channel. The low-pass filtering helps in smoothing frame to frame spectral changes. The rasta functions are used for this purpose. The default RASTA filter parameters are defined using the rastaDefault function. It returns a rastaParamT data structure which contains the necessary parameters needed for the frame-based rasta processing (available from: Feature Extraction).

Table 3.2 lists the feature extraction techniques used with their properties.
Table 3.2 List of Feature extraction techniques with their properties (Santosh et al 2010)

<table>
<thead>
<tr>
<th>Sl.No.</th>
<th>Method</th>
<th>Property</th>
<th>Procedure for Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Principal Component analysis (PCA)</td>
<td>Non linear feature extraction method, Linear map, fast, eigenvector-based</td>
<td>Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data</td>
</tr>
<tr>
<td>2</td>
<td>Linear Discriminate Analysis (LDA)</td>
<td>Non linear feature extraction method, Supervised linear map; fast, Eigen vector-based</td>
<td>Better than PCA for classification</td>
</tr>
<tr>
<td>3</td>
<td>Independent Component Analysis (ICA)</td>
<td>Non linear feature extraction method, Linear map, iterative non-Gaussian</td>
<td>Blind course separation, used for de-mixing non-Gaussian distributed sources(features)</td>
</tr>
<tr>
<td>4</td>
<td>Linear Predictive coding</td>
<td>Static feature extraction method, 10 to 16 lower order coefficient</td>
<td>It is used for feature Extraction at lower Order</td>
</tr>
<tr>
<td>5</td>
<td>Cepstral Analysis</td>
<td>Static feature extraction method, Power spectrum</td>
<td>Used to represent spectral envelope</td>
</tr>
<tr>
<td>6</td>
<td>Mel-frequency scale analysis</td>
<td>Static feature extraction method, Spectral analysis</td>
<td>Spectral analysis is done with a fixed resolution along a Subjective frequency scale i.e. Mel-frequency Scale.</td>
</tr>
<tr>
<td></td>
<td>Filter bank analysis</td>
<td>Filters tuned required frequencies</td>
<td>This method is used to find features</td>
</tr>
<tr>
<td>---</td>
<td>----------------------</td>
<td>------------------------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>8</td>
<td>Mel-frequency cepstrum (MFFCs)</td>
<td>Power spectrum is computed by performing Fourier Analysis</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kernel based feature extraction method</td>
<td>Non linear transformations</td>
<td>Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error</td>
</tr>
<tr>
<td>10</td>
<td>Wavelet</td>
<td>Better time resolution than Fourier Transform</td>
<td>It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform</td>
</tr>
<tr>
<td>11</td>
<td>Dynamic feature extractions i) LPC ii) MFCCs</td>
<td>Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients</td>
<td>It is used by dynamic or runtime Feature</td>
</tr>
<tr>
<td>12</td>
<td>Spectral subtraction</td>
<td>Robust Feature extraction method</td>
<td>It is used on the basis of Spectrogram</td>
</tr>
</tbody>
</table>
Table 3.2 continued

<table>
<thead>
<tr>
<th></th>
<th>Feature</th>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Cepstral mean subtraction</td>
<td>Robust Feature extraction</td>
<td>It is same as MFCC but working on Mean statically parameter</td>
</tr>
<tr>
<td>14</td>
<td>RASTA filtering</td>
<td>For Noisy speech</td>
<td>It is used to find out Feature in Noisy data</td>
</tr>
<tr>
<td>15</td>
<td>Integrated Phoneme subspace/</td>
<td>A transformation based on</td>
<td>Higher Accuracy than the existing Methods</td>
</tr>
<tr>
<td></td>
<td>Compound method</td>
<td>PCA+LDA+ICA</td>
<td></td>
</tr>
</tbody>
</table>

The goal of feature analysis is to extract a set of salient features that characterize the spectral properties of the various speech sounds (the sub word units), which can be efficiently measured. Figure 3.2 gives the block diagram of feature analysis computation.

Figure 3.2 Block diagram of feature analysis computation (adapted from: Rabiner & Juang 2006)
The different features extracted by above techniques are:

- **Spectral feature**: band energies, formants, spectrum and Cepstral coefficients (speaker specific information due to vocal tract)
- **Excitation source feature**: pitch and variation in pitch.
- **Long-term feature**: duration, information energy due to behavior feature (Santosh et al 2010).

Fang et al (2001) said that the performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by (1) the number of filters, (2) the shape of filters, (3) the way that filters are spaced, and (4) the way that the power spectrum is warped. Several experiments are done to find a better implementation. The traditional MFCC calculation excludes the 0th coefficient for the reason that it is regarded as somewhat unreliable. According to the analysis and experiments, the authors find that MFCC(0) can be regarded as the generalized frequency band energy (FBE), which results in the FBE-MFCC. An auto-regressive analysis on the frame energy outperforms its 1st and/or 2nd order differential derivatives. Comparing the traditional MFCC with its corresponding auto-regressive analysis coefficients, the FBE-MFCC and the frame energy with their corresponding auto-regressive analysis coefficients form the best combination, reducing the Chinese syllable error rate (CSER) by about 10.0%. FBE-MFCC with the corresponding auto-regressive analysis coefficients reduces CSER by 2.5%. Experiments are done using Chinese Annotated Spontaneous Speech (CASS) corpus, the FBE-MFCC can reduce the error rate by about 2.9% on an average. Hence depending on the applications, the combinations are chosen. The uniform calculation makes the programming and application of the feature extraction simpler and more straightforward and it can provide an option for different applications.
Ibrahim & Srinivasa (2010) described an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the improvement of speech feature representation in a HMM based recognition approach. An efficient speech recognition system with the integration of MFCC feature with frequency sub band decomposition using sub band coding is used. The two features when passed to the HMM network resulted in better recognition compared to existing MFCC method. The simulation results showed an improvement in the quality metrics of speech recognition with respect to computational time and learning accuracy for a speech recognition system.

3.4.3 Modeling Techniques

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique is divided into two classifications: speaker recognition and speaker identification.

The different modeling techniques used in speech recognition process are discussed below.

3.4.3.1 Acoustic model

Acoustic-phonetic approach has been studied in great depth for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds (Santosh et al 2010).

Christophe et al (2003) presented several solutions for cellular phone embedded speech recognition systems. The proposed techniques are evaluated on a
digit recognition task using both French and English corpora. Three aspects of speech processing are investigated: acoustic parameterization, recognition algorithms and acoustic modeling. Several parameterization algorithms (LPCC, MFCC and PLP) are compared to the Linear Predictive Coding (LPC) included in the GSM norm. The two parameterization algorithms, MFCC and PLP, perform significantly better than the other ones. Moreover, feature vector size can be reduced from 13 to 6 PLP coefficients without a significant loss of performance. Dynamic Time Warping (DTW) and hidden Markov model (HMM) based systems for clean conditions were compared. For HMM systems, an automatic building of phonetic lexicon which increases the system robustness to phoneme substitutions was proposed. The experiments show that HMM systems outperform DTW in speaker independent task. The complexity of this approach is significantly higher than the DTW. Finally, in order to achieve good performance with reasonable resource needs, the HMM model complexity is reduced. Experiments show that reducing the Gaussians per state number from 128 to 1, the increase of WER remains lower than 4%, in the specific context of small vocabulary and isolated word tasks. Tying together all the Gaussian components into a generic speech Gaussian Mixture Model (GMM) reduces the global amount of Gaussian. The phoneme models are derived from this GMM using a well-known MAP (Maximum A Posteriori) adaptation algorithm. The approach based on GMM mapping obtains good results for very compact models (less than 20k parameters). This method was tested on speech-to-text task.

Jordan (2004) deals with the question: “Is Phonetic Embedded Speech Recognition a Disruptive Technology”, discussed about the unused speech recognition feature on the cell phone. These systems allow the user to call 10 to 20 people by name using voice after a training session. The systems tended to fail in noisy environments. The acoustic matching technology used dynamic programming
algorithm, which is well suited to the primitive acoustic models available in the early days of automatic speech recognition, but it was neither effective nor efficient in doing the job at hand – dialing the phone by name.

Zhao & Han (2010) proposed a robust and practical speech recognition system using integrating feature and Hidden Markov Model (HMM) aiming at improving speech recognition rate in noise environmental conditions. A new Genetic Algorithm (GA) for training HMM was proposed. It integrated different speech features into the system, based on global optimization.

The system is comprised of three main sections, a pre-processing section, a feature extracting section and a HMM processing section. Six Chinese vowels were taken as the experimental data. Recognition experiments show that the method is effective and achieves high speed and accuracy for speech recognition. The recognition system has been known to improve speech intelligibility, especially where the acoustic speech signal is degraded by noise. A method for using genetic algorithms to train HMM has been successfully developed. The main contribution of this study is that it presents the idea of searching for the most optimal HMM. The experiments also show that the approach is superior to the classical method. The proposed system was completely simulated on PC. The simulation results show that the approach is correct and effective.

Richard et al (2011) investigated the impact of subspace-based techniques for acoustic modeling in automatic speech recognition (ASR). A new approach to acoustic modeling in ASR, referred to as the subspace based Gaussian mixture model (SGMM), represents phonetic variability as a set of projections applied at the state level in a hidden Markov model (HMM) based acoustic model. The impact of
the SGMM in modeling these intrinsic sources of variability is evaluated for continuous speech recognition (CSR) task. The SGMM is shown to provide an 18% reduction in word error rate (WER) for speaker independent (SI) ASR relative to the continuous density HMM (CDHMM) in the resource management CSR domain. The SI performance obtained from SGMM also represents a 5% reduction in WER relative to subspace based speaker adaptation in an unsupervised speaker adaptation scenario.

George et al (2011) proposed a context-dependent DBN-HMM system that dramatically outperforms strong Gaussian mixture model (GMM)-HMM baselines on a challenging, large vocabulary, spontaneous speech recognition dataset from the Bing mobile voice search task. This system achieves absolute sentence accuracy improvements of 5.8% and 9.2% over GMM-HMMs trained using the minimum phone error rate (MPE) and maximum likelihood (ML) criteria, respectively, which translate to relative error reductions of 16.0% and 23.2%. A novel context-dependent version of DBN-HMMs for LVCSR that achieves substantially better results on the challenging BMVS dataset than strong discriminatively trained GMM-HMMs is described. Although the experiments show that DBN-HMMs provide dramatic improvements in recognition accuracy, training DBN-HMMs is quite expensive compared with training GMMHMMs, primarily because training the former is not easy to parallelize across computers and needs to be carried out on a single GPU machine. However, decoding in DBN-HMMs is very efficient so test time is not an issue in real-world applications. Context-dependent DBN-HMMs is only the first step towards a more powerful acoustic model for LVCSR task.
3.4.3.2 Language model

As explained by Rabiner & Juang (2006), the purpose of the language model or grammar is to provide a task syntax that defines acceptable spoken input sentences. It enables the computation of the probability $P_L(W)$ of the word string $W$.

Nobuo et al (2008) reported two evaluation results of ASR (Automatic Speech Recognition) for a car navigation system. The first interface evaluation results of ASR were using a commercial product and second ASR module evaluation results for noisy in-car speech. In the first evaluation, many undesirable OOV (Out Of Vocabulary) utterances that make the interface worse were reported. To overcome this problem, sophisticated interfaces, which can handle OOV problems, were proposed. ASR module evaluation for the noisy in-car real speech gathered using driving car was carried out. Continuous Speech Recognition (CSR) software Julius/Julian was used.

3.4.4 Pattern Matching

One earlier example of taking advantage of the search redundancy is the dynamic programming method, which turns an otherwise exhaustive search problem into an incremental one. Hence the network that started with 1022 states could be compiled down to a mathematically equivalent network of 108 states that was readily searched for the optimum word string with no loss of performance or word accuracy (Rabiner & Juang 2006).
The pattern-matching approach: It involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well-formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm (Santosh et al 2010).

The job of the pattern-matching module is to combine information (probabilities) from the acoustic model, the language model, and the word lexicon to find the ‘optimal’ word sequence, i.e., the word sequence that is consistent with the language model and that has the highest probability among all possible word sequences in the language i.e., best matches the spectral feature vectors of the input signal (Rabiner & Juang 2006).

Bernhard & Bernhard (1998) describes a prototype implementation of a speech recognition system for embedded applications. The recognition system is comprised of a feature extractor and a classifier. The feature extractor is based on a 64-point Fast Fourier Transformation (FFT); the classifier is based on discrete-density Hidden Markov Models (HMM) with a variable codebook size. Training as well as classification is implemented using the Viterbi algorithm. The prototype is implemented on a digital signal processor (DSP) of type TMS320C40 from Texas Instruments. The recognition rate and the performance are experimentally evaluated using a test vocabulary of 20 words. The recognition is implemented in three consecutive steps: feature extraction, vector quantization and probability calculation (classification). The recognition including these three steps was measured for a typical word of test vocabulary, using a codebook size \( c = 32 \) and a number of states \( N = 5 \). 103 feature vectors were generated for these words that are equivalent to an utterance time of 0.6 s. The total time required to recognize these words is 738 ms. A prototype of an ASR system for command and control applications has been
reported. It allows online recognition with limited memory and runtime. A recognition rate of 99% was achieved by using a test vocabulary of 20 words.

Dmitry (2006) discussed about basics of Automatic Speech Recognition, which include Embedded ASR Systems- Architecture, Optimization, Network Speech Recognition - Transcoded Speech NSR, Bit-Stream NSR, Server Composition, Distributed Speech Recognition – Architecture, Aurora Front-End, Soft Source-Channel Decoding.

Nishanth & Sreenivas (2009) proposed Multi Pattern Viterbi Algorithm (MPVA) to jointly decode and recognize multiple speech patterns for automatic speech recognition (ASR). The MPVA is a generalization of the Viterbi Algorithm (VA) to jointly decode multiple patterns for a given standard Hidden Markov Model (HMM). Unlike Constrained Multi Pattern Viterbi Algorithm (CMPVA), the MPVA does not require the Multi Pattern Dynamic Time Warping (MPDTW) algorithm. The MPVA algorithm has the advantage that it can be extended to connected word recognition (CWR) and continuous speech recognition (CSR) problems. It also gives an improved speech recognition performance over the earlier techniques. Using only two repetitions of noisy speech patterns (-5 dB SNR, 10% burst noise), the word error rate using the MPVA decreases by 28.5 percent, when compared to using individual decoding. MPVA is a generalization of single pattern Viterbi decoding for HMM. A single optimum state sequence is determined for the K set of patterns jointly. The formulation includes the local continuity constraints in determining the optimum path through the (K + 1) dimensional grid. Based on this algorithm, the calculated ASR accuracy is significantly improved over that of single pattern VA. This technique is outperforming CMPVA technique in the presence of noise. The MPVA formulation has the generality of being applicable to many other problems, where robustness of HMM based pattern matching is required.
Revathi & Venkataramani (2011) explored the effectiveness of perceptual features for performing isolated digits and continuous speech recognition. The proposed perceptual features are captured and codebook indices are extracted. Expectation maximization algorithm is used to generate HMM models for the speeches. Speech recognition system is evaluated on clean test speeches and the experimental results reveal the performance of the proposed algorithm in recognizing isolated digits and continuous speeches based on maximum log likelihood value between test features and HMM models for each speech. Performance of these features is tested on speeches randomly chosen from "TI Digits_1", "TI Digits_2" and "TIMIT" databases. This algorithm is tested for VQ and combination of VQ and HMM speech modeling techniques. Perceptual linear predictive cepstrum yields the accuracy of 86% and 93% for speaker independent isolated digit recognition using VQ and combination of VQ & HMM speech models respectively. This feature also gives 99% and 100% accuracy for speaker independent continuous speech recognition by using VQ and the combination of VQ & HMM speech modeling techniques.

Enrico et al (2011) reports on the development and advances in automatic speech recognition for the AT&T Speak4it voice search application. With Speak4it as real-life example, the effectiveness of acoustic model (AM) and language model (LM) estimation (adaptation and training) on relatively small amounts of application field-data is shown. Algorithmic improvements concerning the use of sentence length in LM, of non-contextual features in AM decision trees, and of the Teager energy in the acoustic front-end is introduced. The combination of these algorithms, integrated into the AT&T Watson recognizer, yields substantial accuracy improvements. LM and AM estimation on field-data samples increases the word accuracy from 66.4% to 77.1%, a relative word error reduction of 32%. The
algorithmic improvements increase the accuracy to 79.7%, an additional 11.3% relative error reduction.

**Template based approach matching:** It is unknown speech compared against a set of pre-recorded words (templates) in order to find the best match. This has the advantage of using perfectly accurate word models. Template based approach to speech recognition has provided a family of techniques that have advanced the field considerably during the last six decades. It also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical (Santosh et al 2010).

Jackson & Bruce (2010) proposed a hybrid speech recognition model based on HMM and fuzzy PPM, which has demonstrated to be competitive and promising performance in speech recognition.

**Knowledge based approach:** An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and automatic learning procedure was sought instead. Vector Quantization (VQ) is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. For IWR, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test
speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure.

**Statistical based approach:** The variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM. The approaches represent the current state of the art. The main disadvantage of statistical models is that they must take priori-modeling assumptions, which are answerable to be inaccurate, handicapping the system performance. In recent years, a new approach to the challenging problem of conversational speech recognition has emerged, holding a promise to overcome some fundamental limitations of the conventional hidden Markov model (HMM) approach. This approach is a radical departure from the current HMM-based statistical modeling approaches. For text independent speaker recognition left right HMM is used for identifying the speaker from simple data. HMM has advantages based on Neural Network and Vector Quantization. The HMM is a popular statistical tool for modeling a wide range of time series data. In speech recognition area HMM has been applied to speech classification.

A weighted hidden Markov model HMM algorithm and a subspace projection algorithm are used to address the discrimination and robustness issues for HMM based speech recognition. Word models were constructed for combining phonetic and phonemic models. Learning Vector Quantization (LVQ) method showed an important contribution in producing highly discriminative reference vectors for classifying static patterns. The ML estimation of the parameters via FB algorithm was an inefficient method for estimating the parameters of HMM. To overcome this problem a corrective training method that minimized the number of errors of parameter estimation was developed. A novel approach was used for a
hybrid connectionist HMM speech recognition system based on the use of a Neural Network as a vector quantisation. It showed the important innovations in training the Neural Network. The vector quantization approach showed much of its significance in the reduction of word error rate. MVA method was obtained from modified Maximum Mutual Information (MMI). Various methods are used for estimating a robust output probability distribution (PD) in speech recognition based on the discrete Hidden Markov Model (HMM). An extension of the viterbi algorithm made the second order HMM computationally efficient when compared with the existing viterbi algorithm. A general stochastic model that encompasses most of the models proposed in the literature, pointing out similarities of the models in terms of correlation and parameter time assumptions, and drawing analogies between segment models and HMMs is presented. An alternative VQ method in which the phoneme is treated as a cluster in the speech space and a Gaussian model was estimated for each phoneme. The results showed that the phoneme-based Gaussian modeling vector quantization classifies the speech space more effectively and significant improvements in the performance of the DHMM system have been achieved.

The trajectory folding phenomenon in HMM model is overcome by using continuous density HMM which significantly reduced the word error rate over continuous speech signal. A new hidden Markov model integrating the generalized dynamic feature parameters with model structure was developed. It was evaluated using maximum-likelihood (ML) and minimum-classification-error (MCE) pattern recognition approaches. The authors have designed the loss function for minimizing error rate specifically for the new model, and derived an analytical form of the gradient of the loss function. The K-means algorithm is also used for statistical and clustering algorithm of speech based on the attribute of data. The K in K-means
represents the number of clusters the algorithm should return. As the algorithm starts K points known as cancroids are added to the data space. The K-means algorithm is a way to cluster the training vectors to get feature vectors. In this algorithm the vectors are clustered based on attributes into k partitions. It uses the k means of data generated from Gaussian distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance (Santosh et al 2010).

Rathinavelu & Deng (1996) investigated the interactions of front-end feature extraction and back-end classification techniques in HMM based speech recognizer. The goal was to find the optimal linear transformation of Mel-warped short-time DFT information according to the minimum classification error criterion. These transformations, along with the HMM parameters, were automatically trained using the gradient descent method to minimize measure of overall empirical error count. The discriminatively derived state-dependent transformations on the DFT data were then combined with their first time derivatives to produce a basic feature set. Experimental results showed that Mel-warped DFT features, subject to appropriate transformation in a state-dependent manner, were more effective than the Mel-frequency cepstral coefficients that have dominated current speech recognition technology. The best error rate reduction of 9% is obtained using the new model, tested on a TIMIT phone classification task, relative to conventional HMM. Compared to all three classifiers, THMM produced the lowest error rate and is the new efficient way of utilizing the input data. Mel-warped DFT features, subject to appropriate transformation in a state-dependent manner, are more effective than the MFCCs that have dominated current speech recognition technology.

Jean-Paul (2004) discussed about ASR studied during the past few decades and the systems used in several domains. Most of the present systems are
based on statistical modeling, both at the acoustic and linguistic levels. Noise resistance has become one of the major bottlenecks for practical use of speech recognizers. The models that are presently investigated for increasing recognition performance are presented. The robustness of the systems must be enhanced for the use in adverse conditions like telephone, environmental noise, etc. The on-going efforts toward enhancing the quality of the models used at the acoustic level is presented which will contribute to the development of ASR systems in new application.

Herbordt et al (2005) described the ATRASR large vocabulary speech recognition system developed for ATR. A feature vector consists of 12 MFCCs, and log power is extracted from frames of 20 ms with 10 ms frame shift of data recorded with 16 kHz sampling rate. Cepstral mean subtraction (CMS) is applied. Clean speech Japanese gender-dependent acoustic models are trained using dialogue speech from the ATR travel arrangement task corpus and 25 hours read speech of phonetically balanced sentences. Phoneme-based HMMs with 2086 states generated by the MDL-SSS algorithm with diagonal covariance matrices are used. The system uses a multi-class composite bi gram language model and word-tri gram language models for rescoring. The lexicon size is 55k words. For testing the noise-reduction system, a small database in the cafeteria at ATR was recorded, using the PDA microphone array and a close-talking microphone as reference. Two male speakers and two female speakers read 102 utterances each from the ATR basic travel expression corpus (BTEC) test set- 01. The reverberation time in the cafeteria was about 1 s. The average signal-to-noise ratio (SNR) for each speaker was listed. The frequency range is 50 Hz - 8 kHz. A multi channel speech input device for general purpose PDAs for hands-free speech recognition was presented. The hands-free interface consisted of a real-time implementation of a combination of a robust generalized side lobe canceller and an MMSE estimator for log Mel-spectral energy
coefficients of clean speech. Based on a small experimental database, it was found that both noise-suppression methods have similar performance and that the joint system highly improves the word accuracy of a large vocabulary speech recognizer.

Zaineb & Ahmed (2010) described that evaluating recognition at the phone level is important since the words are always represented by the concatenation of phones units. The behavior of speaker-independent phone recognition in continuous speech based on the technique of HMM was investigated on the selection of an optimal model topology in order to achieve a robust phone recognition system which accomplishes the tradeoff between model size and data training. Correct phone recognition rate of 69.33 percent and accuracy rate of 63.05 was obtained.

Cong-Thanh et al (2010) investigated the recognition of cochlear implant-like spectrally reduced speech (SRS) using Mel frequency cepstral coefficient (MFCC) and hidden Markov model (HMM)-based automatic speech recognition (ASR). The SRS was synthesized from sub band temporal envelopes extracted from original clean test speech, whereas the acoustic models were trained on a different set of original clean speech signals of the same speech database. Changing the bandwidth of the sub band temporal envelopes had no significant effect on the ASR word accuracy. In addition, increasing the number of frequency sub bands of the SRS from 4 to 16 improved the system performance significantly. Furthermore, the ASR word accuracy attained with the original clean speech can be achieved by using the 16, 24, or 32 sub band SRS. The experiments were carried out using the TI-digits speech database and the HTK speech recognition toolkit.

Ahmad et al (2010) discussed about the design and implementation of English digits speech recognition system using Matlab (GUI) based on the Hidden
Markov Model (HMM), which provides a highly reliable way for recognizing speech. The system is able to recognize all English digits from Zero through Nine by translating the speech waveform into a set of feature vectors using Mel Frequency Cepstral Coefficients (MFCC) technique. Two modules called the isolated words speech recognition and the continuous speech recognition were developed. Both modules were tested in both clean and noisy environments and showed a successful recognition rates. In clean environment and isolated words speech recognition module, the multi-speaker mode achieved 99.5% whereas the speaker independent mode achieved 79.5%. In clean environment and continuous speech recognition module, the multi-speaker mode achieved 72.5% whereas the speaker-independent mode achieved 56.25%. However in noisy environment and isolated words speech recognition module, the multi-speaker mode achieved 88% whereas the speaker-independent mode achieved 67%. In noisy environment and continuous speech recognition module, the multi-speaker mode achieved 82.5% whereas the speaker independent mode achieved 76.67%.

Learning based approach: To overcome the disadvantage of the HMMs machine, learning methods could be introduced such as neural networks and genetic algorithm programming. In these machine-learning models explicit rules or other domain expert knowledge need not be given. They can be learned automatically through emulations or evolutionary process (Santosh et al 2010).

Le (1993) explored how a back-propagation neural network (BNN) can be applied to isolated-word speech recognition. Simulation results show that a BNN provides an effective approach for small vocabulary systems. The recognition rate reaches 100% for a 5-word system and 94% for a 10-word system. The general techniques developed can be further extended to other applications, such as sonar target recognition, missile seeking and tracking functions in modern weapon
systems, and classification of underwater acoustic signals. The choice of feature vector plays an important role in the performance of the BNN. The recognition rate may decrease drastically or the system may not converge at all if the features are not correctly chosen. The feature vector chosen in the experiments consisted of the LPC coefficients, short time energy, zero-crossing rate and voiced/unvoiced classification. It worked well and provided good results. However predictions cannot be made about the likely performance of the methods in these areas until they are actually tested.

Raghu et al (1993) describes the development and implementation of a prototype speech recognition system for carrying out isolated word recognition of deaf speech. The recognition system is built around the TMS320C30 DSP processor using a combination of artificial neural networks and conventional signal processing techniques. A vocabulary of 50 words is selected from the modified rhyme test produced by six profoundly deaf persons. The acoustic, temporal and segmental characteristics of their speech were studied to identify features that may be useful in improving the performance of the speech recognition system. These features may account for the high variability and errors. A recognition model was built using artificial neural networks to control the variability, and to use the information concerning the acoustic, temporal and segmental errors in the recognition strategy. The recorded speech samples were randomized and separated into varying numbers of training and testing sets. Selected features were extracted from speech samples and used to train the network to recognize the speech characteristics of each speaker, in a supervised learning fashion. The ability of the network to use the additional features to carry out speaker dependent speech recognition was evaluated. Preliminary results on recognition rates of the developed prototype system indicate that network performance is critically dependent on the number of tokens used in the training phase. The use of additional acoustic features in the recognition task also
improves the recognition rates up to 5%. More features need to be identified that could further increase the performance of recognition systems. The potential of using neural networks in speech recognition tasks has been reviewed and preliminary results indicate that they have the potential to improve the performance of speech recognition systems.

Che et al (1994) explore the use of synergistically integrated systems of microphone arrays and neural networks for robust speech recognition in variable acoustic environments, where the user must not be encumbered by microphone equipment. Existing speech recognizers work best for “high-quality close-talking speech”. Performance of these recognizers is typically degraded by environmental interference and mismatch in training conditions and testing conditions. It is found that the use of microphone arrays and neural network processors can elevate the recognition performance of existing speech recognizers in an adverse acoustic environment, thus avoiding the need to retrain the recognizer, a complex and tedious task. The results showed that a system of microphone arrays and neural networks can achieve a higher word recognition accuracy in an unmatched training/testing condition than that obtained with a retrained speech recognizer using array speech for both training and testing, i.e., a matched training / testing. The system of microphone array and neural network processors can:

• Effectively mitigate environmental acoustic interference.
• Elevate word recognition accuracies of HMM-based and/or DTW-based speech recognizers in variable acoustic environments to levels comparable to those obtained for close-talking, high-quality speech without retraining the recognizer.
• Achieve word recognition accuracies, under unmatched training and testing conditions, that exceed those obtained with a retrained speech recognizer using array speech for both retraining and testing, i.e., under unmatched training and testing conditions.
Flanagan et al (1994) developed a system of microphone arrays (MA) and neural networks (NN) for robust speech recognition. The system expand the power and advantages of existing ARPA speech recognizers to practical acoustic environments where users need not be encumbered by hand-held, body-worn, or tethered microphone equipment, and must have freedom of movement. Examples include Combat Information Centers, large group conferences, and mobile hands-busy eyes-busy maintenance tasks. Use of MA provides auto directive sound pickup that is higher in quality than conventional microphones used at distances. NN processors learn and compensate for environmental interference, and to adapt the testing condition to the training condition. Recognition performance in hostile acoustic environments can thereby be elevated without the need to retrain the recognizer.

Esposito et al (1996) proposed an artificial neural network architecture that detects acoustic features in speech signals and classifies them correctly with English stop consonants [b, d, g, p, t, k] extracted from the general multi-speaker database by modifying some parameter values in the preprocessing algorithm and by using a modified TDNN (Time Delay Neural Network) architecture. Neural networks are accepted as powerful learning tools in pattern recognition in which they proved their performance. Nevertheless, many problems like phoneme classification with multi-speaker continuous speech database are hard even for Neural Networks. The technique performed a good classification which gave the following testing recognition results: 92.9%, 91.8%, 92.4%, 80.3%, 90.2% and 94.2% for b, d, k, p, t & g respectively. It was proved that the Hamming window of 10 msec moved at every 5 msec rate and a sampling rate of 16 kHz are more appropriate when the speech signal is preprocessed using the RASTA-PLP algorithm.
Brett et al (1997) explored the concepts of impulse sampling, Fourier transforms, data windowing, homomorphic filtering, speech coding and classification techniques via MATLAB and NeuralWorks. Applications involving speech coding and phonetic classification were introduced as educational tools for reinforcing signal processing concepts learned in senior level communication classes at the U.S. Naval Academy. These software tools allow sampling an analog speech signal; find the pitch and formant frequencies, and phonetically classifying voice data. The speech coding algorithms used involve digital filtering, data windowing, and spectral analysis. The application provided the means of some of the aspects of diverse signal processing theory in a graphical and procedural manner.

Chee et al (2000) investigated the applicability of artificial neural networks to speech recognition. The synergism of web and phone technologies has led to a new innovative voice web network. The voice web requires a voice recognition and authentication system incorporating a reliable speech recognition technique for secure information access across the Internet. In the experiment, a total number of 200 vowel signals from individuals with different gender and races were recorded. The filtering process was performed using the wavelet approach to de-noise and to compress the speech signals. An artificial neural network, specially the probabilistic neural network (PNN) model, was then employed to recognize and to classify vowel signals into the respective categories. A series of parameter settings for the PNN model in classifying speech signal of vowels was investigated, and the results obtained were analyzed and discussed. Accurate speech recognition requires models that can account for a high degree of variability in the speech signals. The results indicated that the performance of the PNN network was influenced by the smoothing parameter. A small value of smoothing parameter that set the PNN to function as a nearest neighbor classifier yielded the best result.
Abdul et al (2002) discusses an application of ANN to the speech recognition task. The task is to recognize Urdu digits from zero to nine from a mono-speaker and of limited vocabulary train set. A particular class of neural networks called multi layer perceptrons (MLP) that utilize the back propagation of error algorithm for setting of weight is used. After data acquisition, the speech signal is preprocessed and fed to an MLP for classification.

Masoumeh et al (2009) proposed an efficient and effective nonlinear feature domain noise suppression algorithm, motivated by the minimum mean square error (MMSE) optimization criterion. A Multi Layer Perceptron (MLP) neural network in the log spectral domain has been employed to minimize the difference between noisy and clean speech. By using this method, as a preprocessing stage of a speech recognition system, the recognition rate in noisy environments has been improved. The application of the system was extended to different environments with different noises without retraining HMM model. The feature extraction stage was trained with a small portion of noisy data, which was created by artificially adding different types of noises from the NOISEX-92 database to the TIMIT speech database. The proposed method suggests four strategies based on the system capability to identify the noise type and SNR. Experimental results show that the proposed method achieves significant improvement in recognition rates. A new nonlinear noise reduction algorithm motivated by the MMSE criterion in the log spectral domain was developed for environment-robustness speech recognition. The system was developed by using a MLP neural network in the log spectra domain. Experimental results show that this method improves the recognition accuracy in different cases for the TIMIT task and its improvement is greater than that of MBSS. New approach has several key attributes. Only a small portion of the clean speech and the corresponding noisy speech is sufficient for the method to work. The new approach can improve the
recognition accuracy without any extra information about noise such as distribution. It creates a trade-off between system requirements and improvement of recognition accuracy, and knowing the noise type and SNR lead to higher improvement. It is designed to apply to feature extraction outputs and hence can be easily plugged into the feature extraction pipeline of many commonly used ASR systems.

Bo et al (2010) proposed a new speech recognition system design according to the visible characteristics of speech. It is based on multiple neural networks to distinguish different speakers. Pulse Coupled Neural Network (PCNN) was input into the spectrogram for producing the corresponding time series icon as the feature parameters of speech. Then the feature parameters were input into the Probabilistic Neural Networks (PNN) for training PNN to realize speech recognition. The simulation results show higher speech recognition rate if speaker speech signal was extracted by Pulse Coupled Neural Network (PCNN).

Venkateswarlu et al (2011) studied a novel approach for implementing isolated speech recognition. While most of the literature on speech recognition (SR) is based on hidden Markov model (HMM), the present system is implemented by radial basis function type neural network. The two phases of training and testing in a radial basis function type neural network has been described. All classifiers use linear predictive cepstral coefficients. It is found that the performance of radial basis function type neural networks is superior to the other classifier multi layer perceptron neural networks. The promising results obtained through this design show that this new neural networks approach could compete with the traditional speech recognition approaches. Promising results were obtained both in the training and testing phases due to the exploitation of discriminative information with neural networks. It is found that RBF trains and tests faster than MLP. The radial basis
function neural network architecture has been shown to be suitable for the recognition of isolated words. Recognition of the words is carried out in speaker dependent mode. In this mode the tested data presented to the network are same as the trained data. The 16 linear predictive cepstral coefficients with 16 parameters from each frame improve a good feature extraction method for the spoken words, since the first 16 in the cepstrum represent most of the formant information. It is found that the performance of RBF classifier is superior to MLP classifier.

Yusnita et al (2011) discussed speech theories and some methodological concerns about feature extraction and classification techniques widely used in speech recognition system. The isolated word speech recognition is compared with phoneme-based counterpart. Isolated-word ASR for fixed vocabularies was successfully implemented using HMM, ANN and SVM but suffers from lack of adaptability to other languages and increases in complexity as the size of vocabulary increase. Conversely, phonemes, the smallest unit of human speech sounds is apparently more feasible to represent the basic building block for cross-language mapping. The phoneme-based approach has potential to overcome the lack of available training data. It is investigated to achieve a more generic speech recognizer. Isolated word speech recognition has high recognition rate in most of the applications. This depends on the complexity of the system such as speaker independent or dependent, vocabulary size, clean or noisy speech and read or spontaneous word. This type of system suffers from limitation in vocabulary size. As the demand for greater words to be recognized arises, this method is no longer valid and appropriate. Phoneme based system can resolve the problem for unlimited word recognition.

Vikramjit et al (2011) estimated articulatory information in the form of vocal tract constriction variables (abbreviated as TVs) from the Aurora-2 speech
corpus using a neural network based speech-inversion model. Word recognition
tasks were then performed for both noisy and clean speech using articulatory
information in conjunction with traditional acoustic features. The results indicate
that incorporating TVs can significantly improve word recognition rates when used
in conjunction with traditional acoustic features.

The artificial intelligence approach: The artificial intelligence approach
attempts to mechanize the recognition procedure in the same way a person applies
his intelligence in visualizing, analyzing, and finally making a decision on the
measured acoustic features. Expert system is used widely in this approach. The
Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and
pattern recognition approach. In this, it exploits the ideas and concepts of acoustic
phonetic and pattern recognition methods. Knowledge based approach uses the
information regarding linguistic, phonetic and spectrogram. Some speech
researchers developed recognition system that used acoustic phonetic knowledge to
develop classification rules for speech sounds. While template based approaches
have been very effective in the design of a variety of speech recognition systems;
they provided little insight about human speech processing, thereby making error
analysis and knowledge-based system enhancement difficult. On the other hand, a
large body of linguistic and phonetic literature provided insights and understanding
to human speech processing. In its pure form, knowledge engineering design
involves the direct and explicit incorporation of expert’s speech knowledge into a
recognition system. This knowledge is usually derived from careful study of
spectrograms and is incorporated using rules or procedures. Pure knowledge
engineering was also motivated by the interest and research in expert systems.
However, this approach had only limited success, largely due to the difficulty in
quantifying expert knowledge (Santosh et al 2010).
Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms. Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself (Santosh et al 2010).

Min-Lun et al (2006) applied artificial neural network (ANN) to recognize speech. Genetic algorithm (GA) is used to replace the steepest descent method (SDM) for the training of BPNN such that a global search of optimal weight in neural network can improve the performance of speech recognition. The non-specific speaker recognition, which is trained by SDM, the recognition rate to recognize Chinese speech was made with MFCC parameter with recognition rate up to 91%. If BPNN is trained by genetic algorithm, higher recognition, to solve the problem with local optimum, GA was adopted with SDM to improve MSE convergence. Besides increasing GA speed, it also improved system recognition rate up to 95%. Under the condition of adopting only MFCC parameters, speech recognition rate still has room for improvement. Cepstrum coefficient or LPC parameter together with pitch parameter are other modes of calculating features to improve recognition rate.
**Stochastic Approach:** Stochastic modeling entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state Markov model and a set of output distributions. The temporal variability’s are transition parameters in the Markov chain models, while spectral variability’s are the parameters in the output distribution model. These two types of variability’s are the essence of speech recognition. Compared to template-based approach, hidden Markov modeling is more general and has a firmer mathematical foundation (Santosh et al 2010).

Waleed (2002) said that the goal of extraction of the speech segments is to separate acoustic events of interest (the speech segment to be recognized) in a continuously recorded signal from other parts of the background signal. The recognition rate of many voice command systems is very much dependent on speech segment extraction accuracy.

Two novel HMM based techniques that segregate a speech segment from its concurrent background are discussed. The first method can be reliably used in clean environments while the second method, which makes use of the wavelets denoising technique, is effective in noisy environments. These methods have been implemented and shown superiority over other popular techniques, thus, indicating that they have the potential to achieve greater levels of accuracy in speech recognition rates.
3.5 SIMPLE EXAMPLE OF ASR SYSTEM: ISOLATED DIGIT RECOGNITION

Consider a simple isolated word speech recognition system where the vocabulary is the set of 11 digits (‘zero’ to ‘nine’ plus the word ‘oh’ as an alternative for ‘zero’) and the basic recognition unit is a whole word model. For each of the 11 vocabulary words, we must collect a training set with sufficient, say $K$, occurrences of each spoken word so as to be able to train reliable and stable acoustic models (the HMMs) for each word. Typically a value of $K=5$ is sufficient for a speaker-trained system (that is a recognizer that works only for the speech of the speaker who trained the system). For a speaker-independent recognizer, a significantly larger value of $K$ is required to completely characterize the variability in accents, speakers, transducers, environments, etc. For a speaker-independent system based on using only a single transducer (e.g., a telephone line input), and a carefully controlled acoustic environment (low noise), reasonable values of $K$ are on the order of 100–500 for training reliable word models and obtaining good recognition performance. For implementing an isolated-word recognition system, we do the following:

For each word, $n$, in the vocabulary, we build a word-based HMM, $\lambda_n$, i.e., we must (re-) estimate the model parameters $\lambda_n$ that optimize the likelihood of the $K$ training vectors for the $n$-th word. This is the training phase of the system.

For each unknown (newly spoken) test word that is to be recognized, we measure the feature vectors (the observation sequence), $X = [x_1, x_2, \ldots, x_N]$ (where each observation vector, $x_i$ is the set of MFCCs and their first- and second-order derivatives), we calculate model likelihood, $P(X|\lambda_v)$, $1 \leq v \leq V$ for each individual word model (where $V$ is 11 for the digits case), and then we select it as the recognized word whose model likelihood score is highest, i.e.,

$$v = \arg \max P(X|\lambda_v).$$
$1 \leq v \leq V$

This is the testing phase of the system. Figure 3.3 shows a block diagram of a simple HMM-based isolated word recognition system (Rabiner & Juang 2006).

Figure 3.3 HMM-based Isolated word recognizer (adapted from: Rabiner & Juang 2006)

3.6 PERFORMANCE OF SPEECH RECOGNITION SYSTEMS

The performance of speaker recognition system depends on the technique employed in the various stages of speaker recognition system: feature extraction, modeling and testing. The state of art of speaker recognition system mainly used segmentation analysis, Mel frequency Cepstral coefficients (MFCCs), Gaussian Mixture Model (GMM). There are practical issues in the speaker recognition field. Other techniques may also have to be used for a good speaker recognition performance. Some of the practical issues are discussed below.
**Non acoustic sensor** provides an exciting opportunity for multi modal speech processing in areas of application such as speech enhancement and coding. This sensor provides measurement of function of the glottal excitation and can supplement acoustic waveform.

**A Universal Background Model (UBM)** is a model used in a speaker verification system. The model of person independent feature characteristics is compared against a model of person specific feature characteristics to make an accept or reject decision.

**Multi modal** person recognition architecture has been developed for improving overall recognition performance and for addressing channel specific performance. This multi modal architecture includes the fusion of speech recognition system with the MIT/LL and GMM/UBM speaker recognition architecture.

Many powerful speaker recognition systems have introduced high level features, novel classifiers and channel compression methods. SVMs have become a popular and powerful tool in text independent speaker verification. A recent area of significant progress in speaker recognition is the use of high-level features- idiolect, phonetic relations, and prosody. A speaker not only has distinctive acoustic sound but also uses language in a characteristic manner.

A key issue in speech recognition (and understanding) system design is how to evaluate the system’s performance. For simple recognition systems, such as the isolated word recognition system, the performance is simply the word error rate of the system. For more complex speech recognition tasks, such as for dictation applications, we must take into account the three types of errors that can occur in recognition, namely word insertions (recognizing more words than were actually spoken), word substitutions (recognizing an incorrect word in place of the correctly
spoken word), and word deletions (recognizing fewer words than were actually spoken). Based on the criterion of equally weighting all three types of errors, the conventional definition of word error rate for most speech recognition tasks is:

$$\text{WER} = \frac{\text{NI} + \text{NS} + \text{ND}}{\left| W \right|}$$

where NI is the number of word insertions, NS is the number of word substitutions, ND is the number of word deletions, and $\left| W \right|$ is the number of words in the sentence W being scored. Based on the above definition of word error rate, the performance of a range of speech recognition and understanding systems is shown in Table 3.3.

**Table 3.3 Word error rates for a range of speech recognition systems**
(adapted from: Rabiner & Juang 2006)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Type of speech</th>
<th>Vocabulary size</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect digit string (TI database)</td>
<td>Spontaneous</td>
<td>11 (0-9, oh)</td>
<td>0.3%</td>
</tr>
<tr>
<td>Connect digit string (AT&amp;T mall recordings)</td>
<td>Spontaneous</td>
<td>11 (0-9, oh)</td>
<td>2.0%</td>
</tr>
<tr>
<td>Connected digit string (AT&amp;T HMIHY)</td>
<td>Conversational</td>
<td>11 (0-9, oh)</td>
<td>5.0%</td>
</tr>
<tr>
<td>Resource management (RMI)</td>
<td>Read speech</td>
<td>1000</td>
<td>2.0%</td>
</tr>
<tr>
<td>Airline travel information system (ATIS)</td>
<td>Spontaneous</td>
<td>2500</td>
<td>2.5%</td>
</tr>
<tr>
<td>North American business (NAS &amp; WSJ)</td>
<td>Read text</td>
<td>64000</td>
<td>6.6%</td>
</tr>
<tr>
<td>Broadcast news</td>
<td>Narrated news</td>
<td>21000</td>
<td>~15%</td>
</tr>
<tr>
<td>Switchboard</td>
<td>Telephone conversation</td>
<td>45000</td>
<td>~20%</td>
</tr>
<tr>
<td>Call-home</td>
<td>Telephone conversation</td>
<td>28000</td>
<td>~35%</td>
</tr>
</tbody>
</table>

It can be seen that for a small vocabulary (11 digits), the word error rates are very low (0.3%) for a connected digit recognition task in a very clean environment (TI database), but the digit word error rate rises significantly (to 5.0%) for connected digit strings recorded in the context of a conversation as part of a speech understanding system (HMIHY©). Word error rates are fairly low for 1000 to 2500 word vocabulary tasks (RM Linguistic Data Consortium, 1992–2000) and ATIS but increase significantly as the vocabulary size rises (6.6% for a 64,000-word
NAB vocabulary, and 13–17% for a 2,10,000-word broadcast news vocabulary), as well as for more colloquially spoken speech in Switchboard and Call-home, where the word error rates are much higher than comparable tasks where the speech is more formally spoken.

Figure 3.4 illustrates the reduction in word error rate that has been achieved over time for several of the tasks from Table 3.3. It can be seen that there is a steady and systematic decrease in word error rate (shown on a logarithmic scale) over time for every system that has been extensively studied. Hence it is generally believed that virtually any (task-oriented) speech recognition system can achieve arbitrarily low error (over time) if sufficient effort is put into finding appropriate techniques for reducing the word error rate.

Figure 3.4 Reductions in speech recognition word error rates over time for a range of task-oriented systems (adapted from: Rabiner & Juang 2006)
Figure 3.5 gives the comparison of human and machine speech recognition performance for a range of speech recognition tasks.

![Comparison of human and machine speech recognition performance for a range of speech recognition tasks](image)

**Figure 3.5 Comparison of human and machine speech recognition performance for a range of speech recognition tasks** (adapted from: Rabiner & Juang 2006)

If one compares the best ASR performance for machines on any given task with human performance (which often is hard to measure), the resulting comparison shows that humans outperform machines by factors of between 10 and 50; that is the machine achieves word error rates that are larger by factors of 10–50. Hence we still have a long way to go before machines outperform humans on speech recognition tasks. However, one should also note that under a certain condition an automatic speech recognition system could deliver a better service than a human. One such example is the recognition of a long connected digit string, such as a credit card’s 16-digit number, that is uttered all at once; a human listener would not be able to memorize or jot down the spoken string without losing track of all the digits (Rabiner & Juang 2006).
Yifan (1995) presented a survey of 250 publications divided into 3 categories:
- Noise resistance.
- Speech enhancement.
- Model compensation for noise.
Mismatch in training and operating environments as well as mismatch between training and operating conditions were focused. If the properties of noise are known and computing power is cheap then feature-similarity based system is used else transformation-based system is used. Different techniques can be combined. Accurate speech and noise models can be obtained by incorporating a dynamic model (HMM) or frequency correlations like LPC, SOM etc. Weighting portions of speech is based on their SNR. Class dependent processing can be done for different classes like word, phoneme, sound, HMM state and VQ codebook vector.

David et al (2006) said that the availability of real-time continuous speech recognition on mobile and embedded devices has opened up a wide range of research opportunities in human-computer interactive applications. Most of the work in this area to date has been confined to proprietary software, or has focused on limited domains with constrained grammars. A preliminary case study on the porting and optimization of CMU SPHINX-II, a popular open source large vocabulary continuous speech recognition (LVCSR) system, to hand-held devices is presented. The resulting system operates in an average 0.87 times real-time on a 206MHz device, 8.03 times faster than the baseline system. It was claimed that it was the first hand-held LVCSR system available under an open-source license. 142440469X/ 06/$20.00 ©2006 IEEE, a 1000-word vocabulary system operating at under 1 xRT on a 206 MHz hand-held device was presented. Pocketsphinx inherits the easy-to-use API from SPHINX-II, and should be useful to many other developers and researchers in the speech community.
Iman & Manal (2006) developed a speech recognition system (SRS) based on Hidden Markov Models (HMM). The recognition system for Arabic sounds using Artificial Neural Networks was developed. A very good result was obtained compared with those obtained in English phoneme recognition.

Timothy et al (2007) developed the application of two biometric techniques, face and speaker identification for use on mobile devices. It has been found that combining speaker and face identification technologies can have a dramatic effect on person identification performance. In one set of experiments, a 90% reduction in equal error rate in a user verification system was achieved when integrating the face and speaker identification systems. In preliminary experiments examining the use of static and dynamic information extracted from video, it was found that dynamic information about lip movement made during the production of speech could be used to complement information from static lip images in order to improve person identification. Degradation in speaker identification rates in noisy conditions can be mitigated through the use of noise compensation techniques and/or missing feature theory. Noise compensation involves the adjustment of acoustic models of speech to account for the presence of previously unseen noise conditions in the input signal. Missing feature theory provides a multi-modal face and Speaker identification mechanism for ignoring portions of a signal that are so severely corrupted as to become effectively unusable. It was demonstrated that a multi-biometric approach could reduce the equal error rate of a user verification system on a hand-held device by up to 90% when combining audio and visual information. Dynamic information captured from a person's lip movements can be used to discriminate between people, and can provide additional benefits beyond the use of static facial features. The problem of robust speaker identification for hand-held devices was addressed and showed the benefits of the posterior union model and the universal compensation techniques for handling corrupted audio data.
Shing-Tai et al (2007) adopted artificial neural network (ANN) to recognize Mandarin digit speech. Genetic algorithm (GA) was first used to replace Steepest Descent Method (SDM) and make a global search of optimal weight in neural network. The improved GA is then used to train the ANN. We can find that the performance of speech recognition was improved by the later method. The experiment showed that if BPNN is trained by GAs, higher recognition rate is attained. Through out SDM in BPNN speech recognition system, attempting to recognize Chinese speech, the recognition rate up to 91% was achieved. If GA is adopted for training of ANN, the recognition rate can be improved up to 95%. It is shown that the improved GA reveals more excellent learning performance than GA (with two-point crossover) by the experiments. However, the drawback of GA (or improved GA) used to train the ANN is that it will increase training time.

Ashtosh et al (2008) proposed and implemented a robust, low complex voice controller for car audio system involving an efficient Acoustic Echo Cancellation (AEC) and an effective detection module for a simple speech recognizer.

In application such as a voice controlled car audio system, voice commands by the driver are corrupted by audio out of loudspeaker. The proposed method has been implemented on OMAP 2420 TIDSP C55x, with a performance of under 59 Mega Cycles for complete system and is tested in real time. In the car environment the voice commands sent to Car Speech Interface system are corrupted by presence of car audio signal. The system works by using an acoustic echo canceller to cancel the acoustic echo captured by the car Microphone. The AEC uses two adaptive filters to cancel individual audio channels to give better performance. Speech activity detector was used to find the activity regions in Mic signal. A proposed FSD module was used to check false alarms due to residual music
component by exploiting the fact that correlation between music residual and car audio signal is significant. The proposed system showed a best case CER improvement of around 30.3%.

Mangesh & Raghunath (2008) presented a closed-set, text-independent speaker identification using continuous density hidden Markov model (CDHMM). Each registered speaker has a separate HMM which is trained using Baum- Welch algorithm. The system performance has been studied for different system parameters such as the number of states, number of mixture components per state and the amount of data required for training. Identification accuracy of 100% is achieved by conducting the experiments on TIMIT database. The speaker identification rates using an ergodic CDHMM are strongly correlated with the number of states, number of mixtures per state and the amount of data used for training. The performance of HMM lies in having better description of vocal tract and instantaneous characteristic. But the corresponding model needs more speech samples and longer training time. This approach is based on the theory of statistics. The maximum likelihood estimation algorithm used to estimate the parameters, needs a large number of training samples, otherwise the result does not have statistical characteristics.

Astrid & Omid (2009) said that the GPU implementation is about 5.8 times faster than the CPU. Speech recognition is carried out in two phases. In the training phase, the system memorizes a set of reference templates. In the test phase, the system compares a speech signal with all of the reference templates and returns the closest one as the recognized pattern. It should be noted that performance accuracy improves when there are more reference templates to compare with. However, the time to find the closest match increases exponentially with a larger set
of reference templates. This time can be reduced by parallelizing computationally expensive tasks on a graphical processing unit (GPU). There is a marked improvement by moving the computationally expensive tasks onto a GPU.

A few studies attempted to implement a speech recognition algorithm on a graphical processing unit (GPU). In 2007, Dynamic warp algorithm (DTW) was applied to perform voice password identification and was reported that it is possible to obtain an increase in performance by moving the computations onto a graphics card. NVIDIA GeForce 8800 GTX was used in 2008 to compute the acoustic likelihood for their speech recognition system, which is based on a finite-state transducer (FST) framework. The average CPU and GPU times were gathered by computing the acoustic likelihood 2000 times, and reported a performance increase of 33.5% with the GPU implementation. In 2008, the opportunities for parallelizing a more complex speech recognition algorithm was explored. It was a hidden Markov model (HMM) based Viterbi search algorithm, which is typically suited for a large-vocabulary continuous speech recognizer. GPU version proved to be 9 times faster than CPU version.

Yeh-Huann & Raveendran (2009) proposed an HMM-based speech recognition system using adaptive pitch period framing. A hidden Markov models (HMM) based speech recognition system by using cepstrum feature of the signal over adaptive time interval was developed. Pitch period was detected by dyadic wavelet transform and divides the voiced speech signal according to the detected period. System performs HMM-based speech recognition using cepstrum feature to classify the speech signals. Two speech recognition systems have been developed; one is based on constant time framing and the other on adaptive framing. The results are compared and found that adaptive framing method shows better result in both
data distribution and recognition rate. This study was compared with fixed frames and the results showed the proposed method produced better recognition rate even in the presence of noise.

Zheng-Hua & Borge (2010) discussed a low-complexity and effective frame selection approach based on a posteriori signal-to-noise ratio (SNR) weighted energy distance: The use of an energy distance, instead of, e.g., a standard cepstral distance, makes the approach computationally efficient and enables fine granularity search, and the use of a posteriori SNR weighting emphasizes the reliable regions in noisy speech signals. It is experimentally found that the approach is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. The resulting variable frame rate analysis method is applied to three speech-processing tasks that are essential to natural interaction with intelligent environments. First, it is used for improving speech recognition performance in noisy environments. Second, the method is used for scalable source coding schemes in distributed speech recognition where the target bit rate is met by adjusting the frame rate. Third, it is applied to voice activity detection. Very encouraging results are obtained for all three speech-processing tasks.

Tobias et al (2010) presented a technique for fast adaptation of speech and speaker related information. Fast learning is particularly useful for automatic personalization of speech-controlled devices. Such a personalization of human-computer interfaces to be used in intelligent environments represents an important research issue. Speech recognition is enhanced by speaker specific profiles, which are continuously adapted. A fast but robust tracking of speaker characteristics and optimal long-term adaptation are investigated to avoid an extensive enrollment of
new speakers. Implementation suitable for speaker specific speech recognition in adverse intelligent environments was presented. Exemplarily, in-car applications such as speech controlled navigation; hands-free telephony or infotainment systems are investigated for embedded systems. Results for a subset of the SPEECON database are presented. They validate the benefit of the presented speaker adaptation scheme for speech recognition. Speaker characteristics are captured after very few utterances. In the long run speaker characteristics are accurately represented. This adaptation scheme might be used to develop an unsupervised speech controlled system comprising speech recognition and speaker identification. The experiments have shown that 25% relative improvement compared to the speaker independent speech recognizer can be achieved with respect to the error rate.


The major issues in speaker recognition for mobile devices are (i) presence of varying background environment, (ii) effect of speech coding introduced by the mobile device, and (iii) impairments due to wireless channel. Multi-SNR multi-environment speaker models and speech enhancement (preprocessing) methods for improving the performance of speaker recognition system in mobile environment are proposed. Five different background environments (Car, Factory, High frequency, pink noise and white Gaussian noise) have been simulated using NOISEX data. Speaker recognition studies were carried out on TIMIT, cellular, and microphone speech databases. Auto associative neural network models are explored for developing these multi-SNR multi-environment speaker models. The results indicate that the methods have enhanced the speaker recognition performance in the presence of different noisy environments.
Urmila & Vilas (2010) proposed many feature extraction algorithms developed for noisy environment that are designed specifically to have a low sensitivity to background noise. Some features extraction techniques with their pros and cons are discussed. Some new methods are developed using combination of more techniques. Study confirms that, new developed techniques give better performance in noisy environment where MFCC fails. There is a need to develop some more new hybrid methods that will give improved performance in robust speech recognition in noisy environment.

Xiaoqiang et al (2010) used information retrieval (IR) techniques to improve a speech recognition (ASR) system. The potential benefits include improved speed, accuracy, and scalability. Where conventional HMM-based speech recognition systems decode words directly, IR-based system first decodes sub word units. These are then mapped to a target word by the IR system. In this decoupled system, the IR serves as a lightweight, data-driven pronunciation model. The proposed method is evaluated in the Windows Live Search for Mobile (WLS4M) task, and best system has 12% fewer errors than a comparable HMM classifier. Using an inexpensive IR weighting scheme (TF-IDF) yields a 3% relative error rate reduction while maintaining all of the advantages of the IR approach. The integration of information retrieval techniques in this system provides a flexible pronunciation modeling for the ASR. The proposed system is able to outperform a traditional HMM based ASR system.

Sungwoong et al (2010) considers a large margin training of semi-Markov model (SMM) for phonetic recognition. The SMM framework is better suited for phonetic recognition than the hidden Markov model (HMM) framework. The SMM
framework is capable of simultaneously segmenting the uttered speech into phones and labeling the segment-based features. The SMM framework is used to define a discriminant function that is linear in the joint feature map, which attempts to capture the long-range statistical dependencies within a segment and between adjacent segments of variable length. The parameters of the discriminant function are estimated by a large margin-learning criterion for structured prediction. The parameter estimation problem, which is an optimization problem with many margin constraints, is solved by using a stochastic sub gradient descent algorithm. The proposed large margin SMM outperforms the large margin HMM on the TIMIT corpus. The LMSMM for phonetic recognition was proposed. Under the SMM framework, a linear discriminant function and segment-based joint feature map which consist of the transition feature, duration feature, and content feature are defined. The function parameters were estimated by the large margin training based on the SSVM and the stochastic sub gradient descent algorithm. Experimental results showed that the proposed LMSMM outperformed the LMHMM on the TIMIT phonetic recognition.

Wei-Tyng (2010) made a study of applying Hidden Conditional Random Fields (HCRF) to establish speaker models. The results confirm that the HCRF has good capabilities on speaker identification. A novel training technique combining discriminative training algorithm on HCRF model is proposed. This study also adopted discriminative training technique to train GMM, HMM, and HCRF speaker models respectively; and we investigated the performance of speaker identification with the three speaker models with different amounts of training speech for clean and noisy testing speech. The experimental results indicate that the HCRF model consistently achieved the lowest error rate among the three models regardless of the length of the test and training speech and presence of noise. The best performance was achieved by the HCRF scheme for MATDB-4 testing speech under 20
enrolment utterances per speaker, which led to 23.0% and 15.4% decreases in error rate, respectively, compared with the results by the GMM and HMM schemes.

Zhou & Zheng (2010) proposed a multi-dimensional time series data mining model for the meteorological data. In this model extraction, difficulty of data mining is reduced. Kmeans cluster is used to make the symbols of sequence. Rule extraction is used for getting useful rules in experiments. The results of experiment show that this model has a great practicability. Based on the background of analyzing and applying the ground-based automatic weather station data, by the study of picking dimension, piecewise linear fitting method, the symbols of cluster sequence, the dimensions redundant reduction algorithm is proposed.

Ibrahim & Srinivasa (2010) presented an approach to the recognition of speech signal using frequency spectral information with mel frequency for the improvement of speech feature representation in a HMM based recognition approach. The mel frequency approach exploits the frequency observation for speech signal in a given resolution, which results in resolution feature overlapping resulting in recognition limit. The simulation results show an improvement in the quality metrics of speech recognition with respect to computational time and learning accuracy for a speech recognition system. A speech recognition system for speech recognition at noise corruption is developed. The MFCC algorithm, which cannot extract the feature of speech signal at lower frequency, is modified. An efficient speech recognition system with the integration of MFCC feature with frequency sub band decomposition using sub band coding is proposed. The two features passed to the HMM network result in better recognition compared to existing MFCC method. It is observed that the implemented system gives better efficiency compared to the existing method.
Rajesh et al (2010) developed a speaker identification system, which is used to determine the identity of an unknown speaker among several speakers of known speech characteristics, from a sample of his or her voice.

Every speaker has different individual characteristics embedded in his/her speech utterances. These characteristics can be extracted from utterances and different neural network models are used to get the desired results. To evaluate speech characteristics from utterances they are stored in digitized form. Speech features namely LPC, RC, APSD, number of zero crossing and formant frequencies are extracted from speech signal and formed speech feature vectors. These data features are fed into artificial neural network using back propagation learning algorithm and clustering algorithm for training and identification processes of different speakers. The database used for this system consists of 20 speakers including both male and female from different parts of India and languages are Hindi, Sanskrit, Punjabi and Telugu. The average identification rate of 83.29% is achieved when the network is trained using back propagation algorithm and it is improved by about 9% and reached up to 92.78% when using clustering algorithm.

Tobias et al (2010) presented a self-learning speech controlled system based on a unified modeling of speech and speaker characteristics. Multiple recognitions are avoided because speaker identification is computed in parallel to speech recognition. Computation on different time scales was introduced. Speaker specific speech recognition is realized by an online codebook selection. On an utterance level a more reliable guess of the speaker identity is calculated in parallel to speech recognition. The identification result is employed for continuous speaker adaptation. By combining EV and ML estimates in the MAP framework, recurring
speakers can be tracked in an unsupervised manner after a short enrolment of two utterances. It is evaluated that WA could be significantly increased compared to the speaker independent baseline or short-term adaptation without speaker tracking. 94.64% speaker identification rate and 88.20% speech recognition rate were achieved in the experiments carried out. The results of the baseline and the corresponding upper bound with pre-defined speaker identity were 85.23% and 88.90% WA.

Raji et al (2010) discussed a novel technique for recognition of the isolated question words from Malayalam speech query. A database consisting of 250 isolated question words was created and analyzed. Discrete Wavelet Transform (DWT) is used for the feature extraction purpose and Artificial Neural Network (ANN) is used for classification and recognition. A recognition accuracy of 80% could be achieved from this experiment. An overall recognition accuracy of 72% for five question words was achieved.

Raji et al (2010) discussed a novel technique for recognition of the isolated question words from Malayalam speech query. A database consisting of 500 isolated question words was created and analyzed. Fast Fourier Transform (FFT) and Discrete Cosine transform (DCT) was used for the feature extraction purpose and Artificial Neural Network (ANN) was used for classification and recognition. A recognition accuracy of 85% could be achieved from this experiment. Artificial neutral networks can be used for reduced computational complexity. The original input signal through the inverse DCT could be successfully reconstructed. We used Multi Layer Perceptron (MLP) architecture was used for the training and testing of input feature vectors. A Neural Network in 32 bit C language was developed based
on the Back Propagation Algorithm. Neural classifier can handle input feature vector dimensions up to 100, with 100 hidden layers. The output nodes in the developed classifier correspond to the actual number of predicted outputs. Sigmoid was used as the activation function for the inputs. A database proportion of 80:20 was used for training and testing of the classifier. The feature vectors obtained in this experiment have a dimension of 50 coefficients for each spoken digits. These feature vector parameters are given to the input values for training the Neural Network. The obtained weights from the training phase are used to test the testing vectors. An overall accuracy of 85% has been achieved from this experiment.

Shi-Huang et al (2011) described a speech-controlled interface with cloud computing technology for vehicle on-board diagnostic (OBD) system. The vehicle OBD system is constructed of embedded global position system (GPS-OBD) module and vehicle surveillance server. The speech recognition task is performed in vehicle surveillance server, instead of GPS-OBD module. The speech signal spoken by driver is transferred into AMR format and then submitted to the speech recognition system built in the vehicle surveillance server via 3.5G wireless network. The speech features are extracted from AMR bit stream and carries out accurate speech recognition results. Simultaneously, the vehicle equipped with GPS-OBD module will consistently report its OBD-II data and GPS coordinates to the vehicle surveillance server. The main advantage of the system is that users do not need to install any speech recognition software in their GPS-OBD module before using the cloud speech controlled interface. This will also benefit the cost and size reduction of the GPS-OBD module. The driver can acquire the current vehicle status through speech commands.
Jungpyo et al (2011) proposed a novel speech enhancement algorithm. The algorithm controls the amount of noise reduction according to whether the target speech is absent or present in noisy environments. Based on the estimated speech absence probability (SAP), the amount of noise reduction is adaptively controlled. To calculate the SAP, normalized cross correlation of linear predictive residual signals is utilized instead of original input residual signals. It is especially robust and effective in reverberant and realistic environments. Experimental results show that the proposed algorithm improves speech recognition rates compared with conventional beam forming algorithms in car environments. The ALRC based on a probabilistic approach is proposed. Experimental results show considerable performance improvements in SNR and SRR. Compared with the conventional LCMV, the proposed algorithm achieved sufficient background noise reduction and target signal preservation. Also, the noise reduction performance of the proposed algorithm was better than that of the state-of- the-art noise reduction algorithms. The LCMV with the ALRC is a novel speech enhancement technique compensating for the weakness of the LCMV, indiscriminate noise reduction regardless of speech presence.

Chapter 4 considers in detail some selected techniques discussed by different authors.