1.1 INTRODUCTION

Speech is the most natural and effective mode of communication. With the rapid development of the information technology and increasing popularity of digital computer, human-machine communication using natural speech has received wide attention from both academic and business communities. Developing a system that can understand natural language has been a continuing goal of Artificial Intelligence researchers. The term "natural" languages refer to the languages that people speak, like Assamese, Bodo, English etc., as opposed to artificial languages like programming languages or logic. One of the most critical test for intelligent behaviour of a system as proposed by Alan Turing in 1950 is it's the ability to communicate with their human counterpart in a natural way. The natural language is one of the most important medium for such a communication.

We say that a system understand a natural language if it behaves by taking a predictably correct and acceptable action in response to the natural language input.

Developing a program that can understand natural language is a difficult problem. Natural languages are large - they contain infinitely many sentences. No matter how many sentences a person has heard or seen, new one can always be produced. Also, there is much ambiguity in a natural language. Many words can have several meaning. Moreover, the same word or sentence may have different meaning in different contexts. These features make it difficult to design a system that can understand natural language.
The basic difference between a system that can understand natural language and an ordinary data processing system is that the former can process written and spoken human language as language. To understand the meaning of a natural language, the system must have the linguistic knowledge of that language. A language understanding program must have considerable knowledge about the structure of the language including what the words are and how they are combined into phrases and sentences. It must know the meaning of the sentences and the context in which they are being used. Finally, the system must have some general world knowledge as well as knowledge of what human knows and how they reason.

The linguistic knowledge need for understanding the natural language can be classified into the following categories –

**Phonological:** The phonological knowledge relates sounds to the words we recognize. A phoneme is the smallest unit of sound. Phones are aggregated into word sound.

**Morphological:** Morphological knowledge is the lexical knowledge that relates to word constructions from basic units called morphemes. A morpheme is the smallest unit of meaning, for example, the construction of the word *nicely* from root *nice* and suffix *ly*.

**Syntactic:** The syntactic knowledge relates how words are put together or structured to form grammatically correct sentences in a language.

**Semantic:** The semantic knowledge is concerned with the meaning of words and phrases and how they combine to form meaning of the sentences.

**Pragmatic:** The pragmatic knowledge is the high-level knowledge relating the use of sentences in different contexts and how the context affects the meaning of the sentences.
World: The world "knowledge" is the knowledge related to the language which an user must posses in order to understand and carry on a conversation. It must include an understanding of the other person's beliefs and goals.

When a string of word has been detected, the sentences are parsed or analysed to determine their structure (syntax) and grammatical correctness. The meanings (semantics) of the sentences are then determined and appropriate representation structures created for the inferencing programs. The whole process consisted of a series of transformations.

1.2 A BRIEF HISTORY OF SPEECH RECOGNITION RESEARCH

The earliest attempts in the field of Automatic Speech Recognition were reported in 1950's. In 1952, at Bell Laboratories, Davis, Biddulph and Balashek [29] built a system for isolated digit recognition for single speaker. In 1956, at RCA Laboratories, Olson and Belar [88] developed a system designed to recognize 10 distinct syllables of a single speaker. In 1959, at University College in England, Fried and Dener [37] demonstrated a system designed to recognize four vowels and nine consonants. In the same year, at MIT's Lincoln Laboratories, Forgie and Forgie [35] built a system to recognize 10 vowels in a speaker independent manner. All these systems used spectral information to extract voice features.

time aligning a pair of speech utterances. This work remained unknown until the early 1980's. A major achievement of 1960's was the pioneering research of Reddy [100] in continuous speech recognition by dynamic tracking of phonemes.

In the 1970’s researchers achieved a number of significant milestones, mainly focusing on isolated word recognition. This effort made isolated word recognition a viable and usable technology. Itakura's research [46] showed how linear predictive coding can be applied in speech recognition tasks. Sakoe and Chiba [107] showed how to apply dynamic programming. Velichko and Zagoruyka [121] reported the use of pattern recognition technique in speech understanding. During 1970’s IBM had made remarkable contribution in the field of speech recognition research specially in the development of large vocabulary speech recognition system [47, 48]. Researcher at AT&T Bell Laboratories made a series of experiments on speaker independent speech recognition techniques [98].

Towards the end of 1980’s, the entire speech related technology was mostly confined on the recognition of speech. Along with the advent of speech research, a remarkable change in the shift of technology from template based approach to statistical method – i.e., the Hidden Markov Models (HMM) [33, 97] was entertained. Another idea that appeared in the arena was the use of neural network in the speech recognition problem [71, 125].

From the 1990’s the speech recognition research take a new dimension. It was not only concerned with recognizing the word contents but also prosody [5, 6, 7, 8, 74, 105] and personal signature [24, 26, 76, 102, 103].

Despite all the advantages in the speech recognition area, the problem is far from being completely solved. A number of excellent commercial products, which are getting closer and closer to the final goal, are currently available in the market.
Although these applications are seemed impressive but are computationally intensive. So, in order to make their usage widespread more effective algorithms must be developed.

During the last few years, much effort has been made by researchers’ to explore the applicability of neural network in various speech recognition tasks [2, 17, 19]. Indeed, neural networks have been found to possess a number of inherent properties that make them specially attractive for speech recognition. This includes: (1) Neural network having good learning capability which is essential for automatic characterization of speech variability, (2) Neural networks learning algorithms required no explicit assumption on the statistical properties of speech data, (3) Neural network has powerful generalization capability which is desirable when only a limited amount of training data are available and (4) Neural network have parallel architecture that can handle the high computational requirements for speech processing.

1.3. INTRODUCTION TO AUTOMATIC SPEECH RECOGNITION SYSTEM

During the last many decades attempts were made by many researchers to built an "intelligent machine" which can master the natural speech. In its simplest form, this machine should consist of two subsystems, namely automatic speech recognition (ASR) and speech understanding (SU). The goal of automatic speech recognition is to transcribe natural speech, while speech understanding is to understand the meaning of the transcription. Recognising and understanding a spoken sentence is obviously a knowledge-intensive process which must take into account all variable information about the speech communication processes, from acoustics to semantics and pragmatics.
The designing of an ASR system is related to a large variety of disciplines such as acoustics, signal processing, pattern recognition, phonetics, linguistics, psychology and neuroscience etc. ASR is a complicated task. Although many advancements have been made in ASR research, the performance of today's best systems is more than an order of magnitude in error rate from human performance. The difficulties can be best described in terms of the tasks to be performed. These include:

**Number of speakers:** With more than one speaker, an ASR system must cope with the difficult problem of speech variability from one speaker to another. This is usually achieved through the use of large speech database as training data.

**Nature of the utterance:** *Isolated word recognition* restricts the speaker to insert artificial pause between successive utterances. *Continuous speech recognition* systems are able to cope with natural speech utterances in which words may be tied together and may, at times be strongly affected by co-articulation. *Spontaneous speech recognition* systems allows the possibility of pause and false starts in the utterance, the use of words not found in the lexicon, etc.

**Vocabulary size:** In general, increasing the size of the vocabulary decrease the recognition scores.

**Language complexity:** The task of continuous speech recognisers is simplified by limiting the number of possible utterances through the imposition of syntactic and semantic constraints.

**Environment conditions:** The adverse conditions (such as noise, distorted signal, and transmission line variability) of a site can drastically degrade the system performance.

There are three major approaches of speech recognition [95]. First, the *acoustic-phonetic approach*, where it is assumed that the phonetic units are broadly
characterised by a set of features, such as formant frequency, voiced/unvoiced, and
pitch. These features are extracted from the speech signal and are used to segment and
label the speech

Second, the **pattern recognition approach** requires no explicit knowledge of
speech. This approach has two steps - namely, training of speech patterns based on
some generic spectral parameter set and recognition of patterns via pattern
comparison. The popular pattern recognition techniques include template matching,
Hidden Markov model (HMM), and artificial neural network (ANN)

Third, the **artificial intelligence approach** attempts to mechanise the
recognition procedure according to the way a person applies its intelligence in
visualizing, analysing, and finally making a decision on the measured acoustic
features. Expert systems are used widely in this approach.

### 1.4 CONVENTIONAL PATTERN RECOGNITION APPROACHES FOR
SPEECH RECOGNITION

Traditionally, Speech Recognition was performed by template matching
method [45, 98]. These methods involved direct comparison between parametric
feature representations of input speech signal and a set of pre-stored reference
templates. Since speech is generally represented as time sequence of spectral feature
vectors, global time alignment and normalization are essential for utterances of
different length. To solve this problem, dynamic time wrapping (DTW) was
developed. The algorithm was based on dynamic programming [12]. The DTW
algorithms are well formulated and easy to implement. They have been widely used in
many practical applications of isolated word recognition [82,85,107].
In the template matching method, the reference templates are created via a preliminary training process. This is done by applying some kind of averaging procedure over a large collection of training data. In the simplest case, a single template is used to represent each word and the recognition result is produced by identifying the closest word template. For more sophisticated applications, multiple reference templates are usually required to represent different physical realization of the word. The recognition decision is then made k-nearest neighbour (kNN) algorithm [31,95].

However, template matching is inherently not aimed at characterizing the speech signal in a strictly statistical sense. It is often inadequate to handle large speech variabilities, particularly in speaker independent cases[95]. Furthermore, for large vocabulary application, the number of reference template required would be prohibitively excessive in terms of both storage and computation. To overcome these limitations, a number of statistical modelling techniques were proposed[10,46,48], among which the well established Hidden Markov Model (HMM) approaches was predominating over the others for the past few decades [51,63]. In an HMM, speech signal is assumed to be a combination of two concurrent stochastic processes. On one hand, the template structure of speech is modelled as state transitions in a Markov chain. On the other hand, the locally stationary spectral properties are described by a set of probability density functions (PDFs) governing individual states. These states are regarded as being "hidden" in the sense that, at any particular instant of time, their occurrences are not determined but can only be estimated based on the given spectral features.

The main strength of HMMs can be attributed to the strong mathematical basis for both training and recognition. The training of an HMM involves simultaneous
determination of the state transition probability \((a_{ij})\) and the PDF \(b_j(x)\). This is done via an efficient iterative procedure referred as the Baum-Welch re-estimation \([10, 97]\), which is aimed at likelihood maximization (ML) for all training utterances. For isolated word recognition, each word is represented by a separate HMM. The \textit{a posteriori probability} of input utterance is computed for each HMM using the so-called \textit{forward-backward} evaluation procedure \([97]\). The model with maximum probability will then indicate the recognized word.

The HMM methods have been studied extensively with successful applications in a variety of practical systems \([3, 17, 92, 140]\). Indeed, they are still regarded as the most powerful speech recognition technique. Nevertheless, the standard HMM formalism is also exposed to a number of limitations \([17, 96]\). First of all, individual acoustic vectors are treated in HMM as independent of each other. Obviously this is a fallacious perspective since speech is produced by a set of continuously moving apparatus instead of by mixing up the output of many uncorrelated signal generators. The current states of these apparatus will certainly affect their future states. Second, the maximum likelihood training criterion does not necessarily lead to the minimization of recognition error rate. Most effective discriminative training algorithms are then needed to attain high recognition accuracy. Third, HMM accept short-time acoustic features in the present HMM framework. Lastly, the model topology and statistical distribution for each state are selected beforehand are not flexible to suit unknown variabilities of training data.

The new emerging technology falls into the realm of artificial neural networks (ANNs) \([17,36,72,172]\). Unlike HMMs, the most important development in ANNs or connectionist models was not driven by their speech applications. During the past few decades, much effort has been devoted by researchers to explore the applicability of
neural network in various speech recognition tasks [23, 136]. Indeed, neural network have been found to possess a number of inherent properties that make them specially attractive for speech recognition [17, 95]. These include:

a) Neural network have good learning capability, which is absolutely essential for automatic characterization of speech variability

b) Neural network learning algorithms required no explicit assumption on the statistical properties of speech data.

c) Neural network have powerful generalization capability which is desired when only a limited amount of training data are available.

d) Neural network has parallel architectures that can handle the high computation requirements for speech processing.

The most conventional neural networks used as static pattern classifiers only, e.g., multi-layer perceptron (MLP), self-organized map (SOM) and learning vector quantization (LVQ), are not suitable for sequentially arranged speech patterns. Thus, exploration of the temporal processing capability in neural network is a challenging task [21].

1.5 INTRODUCTION TO NEURAL NETWORK

It has been a long standing technical challenge to let machine acquire and expand certain human abilities. Apart from the higher level functions of the brain, the most difficult tasks are speech and image recognition. Modern computers outperforms the human brain by orders of magnitude as far as execution speed of the elementary operations are concerned, but it is far behind when it comes to deal with more complex perceptive tasks. It is therefore quite natural that scientists and engineers are inspired by biology to build more elaborate machines. One approach is to imitate the
human brain by Artificial Neural Network (ANN), though ANNs are only a very poor imitation of the human brain. Artificial Neural Network, which is simplified model of the biological neuron system, is a massively parallel distributed processing system made up of highly interconnected neural computing elements that have the ability to learn and thereby acquiring knowledge.

1.5.1 BIOLOGICAL NEURAL NETWORK

The functionality of human brain is yet to be completely understood. However, the concept of neuron as the fundamental constituent of the brain attributed by Romón Y. Cajal (1911), has made the study of its functioning comparatively easier.

Human brain contains about $10^{10}$ basic units called neuron or nerve cell. Each neuron, in turn, connected to about $10^4$ other neurons. A neuron is a small cell that receives electro-chemical signals from its various sources and in turn responds by transmitting electrical impulses to other neurons. An average brain weights about 1.5 kg and an average neuron weights about $1.5 \times 10^{-9}$ grams. Some of the neurons perform input and output operations, referred to as *afferent* and *efferent* cells, the remaining fraction forms as part of interconnected network of neurons which are responsible for signal transformation and storage of information. Despite their different activities, all neurons share common characteristics.

A neuron is composed of a nucleus – a cell body known as *soma*. One end of the cell, the input end, has a number of fine processes called *dendrites* because of their resemblance to a tree (dendro is a Greek root means tree).

Most neurons have a long, thin process, the *axon* that leaves the cell body and may run for a meter. The axon is the transmission line of the neuron. Axons can give
raise to collateral branches, along with the main branch, so the actual connectivity of the neuron can be quite complicated. Neurons are amongst the largest cells in the human body and are certainly the most extended. When the axons reach their final destination they branch again, which is called terminal arborization (arbor in Latin root means tree). The ends of the axonal branches are complex, highly specialized structures and are called synapses.

In a standard neuron, dendrites receive inputs from other cells, the soma and dendrites process and integrate the inputs and information is transmitted along the axon to the synapses, whose output provide input to other neurons or the effectors organ.

The transmission of signal from one cell to another at a synapse is a complex chemical process in which specific transmitter substances are released from the sending side of the junction. The effect is to raise or lower the electrical potential inside the body of the receiving cell. If this potential reaches a threshold, an electrical activity in the form of short pulses is generated. When this happens, the cell is said to have fired. These electrical signals of fixed strength and duration are sent down the axon. Generally, electrical activity is confined to the interior of a neuron, whereas the chemical mechanism operates at the synapses.

In the state of inactivity, the interior of the neuron - the protoplasm is negatively charged against the surrounding neural liquid containing positive sodium (Na⁺) ions. The resulting resting potential of about -70 mV is supplied by the action of the cell membrane, which is impenetrable for the positive sodium ions. This causes a deficiency of positive ions in the protoplasm. Signals arriving from the synaptic connections may result in a temporary depolarization of the resting potential. When the potential is increased to a level above -60 mV, the membrane suddenly loses its
impermeability against Na\(^+\) ions, which enter into the protoplasm and reduce the potential difference across the membrane. The sudden change in the membrane potential causes the neuron to discharge. When this happens, the neuron is said to be fired. The membrane then gradually recovers its original properties and regenerates the resting potential over a period of several milliseconds. During the recovery period, the neuron remains incapable of further excitation. The discharge which initially occurs in the cell body propagates as signal along the axon to the synapses. The discharge signal travelling along the axon stops at the synapses, as there is no conducting link to the next neuron. Transmission of signal across the synaptic gap is mostly affected by chemical activity. When the signal arrives at the pre-synaptic nerve terminal, special substances called neurotransmitters are produced in tiny amounts. The neuron transmitted molecules which travel across the synaptic junction and reaching the postsynaptic neuron within about 0.5 ms. These substances modify the conductance of the postsynaptic membrane for certain ions, causing a polarization or depolarization of the postsynaptic potential. If the induced polarization potential is positive, the synapse is termed as excitatory, because the influence of the synapse tends to activate the postsynaptic neuron. If the polarization potential is negative, the synapse is called inhibitory, since it counteracts the excitation of the neuron. All the synaptic ending of an axon are either of an excitatory or inhibitory.

The cell body of the neuron acts as a summing device due to the net depolarizing effect of its input signals. The net effect decays with a time constant of about 5-10 ms. But if several signals arrive within such a period, their excitatory effect accumulate. When the total magnitude of the depolarization potential of cell body exceeds the critical threshold (about 10 mV), the neuron fires.
The complexity of the human central nervous system is due to the large number of neurons and their mutual interconnections. Connectivity is characterized by the complementary property of convergence and divergence of the nervous system. In the human cortex every neuron is estimated to receive a converging input on an average from about $10^4$ synapses. On the other hand, each cell feeds its output into many hundreds of other neurons. The total number of neuron in the human cortex is estimated to be in the vicinity of $10^{10}$ Thus, combining with the average number of synapses per neuron, this yields a total of about $10^{14}$ synaptic connection in the human brain.

A comparison of the functionality of the neural network with that of a modern computer is made with respect to the following five characteristics:

1) **Speed:** - Neural networks are slow in processing information. For the most advanced computers the cycle time corresponding to execution of one step of a program in the central processing unit is in the range of few nanoseconds where as the cycle time for a neuron event prompted by the external stimulus occurs in millisecond range. Thus the computer processes data nearly a million times faster than the brain.

2) **Processing:** - The real power of brain comes from its massively parallel processing capability. Parallel processing is one where more than one instruction can be executed simultaneously. Most programs have large number of instructions, and they operate in a sequential mode, one instruction after another on a conventional computer. On the other hand, the brain operates with massively parallel operations, each of them having comparatively fewer steps. It explains the superiority of human brain over computer in information processing for certain task, despite being several order of magnitude slower compared to the computer in processing information.
3) **Size and Complexity**: - Neural network has larger number of computing elements, and the computing is not restricted to within neurons. The number of neurons in a brain is estimated to be about $10^{10}$ and the total number of interconnection to be around $10^{14}$. Thus, the size and complexity of connections enabled the brain with a power of performing complex pattern recognition tasks which are not possible to realize in today's computer. The complexity of brain is further compounded by the fact that computing takes place not only inside the cell body or soma but also outside in the dendrites or synapses.

4) **Storage**: - Neural networks store information in the strengths of the interconnections. In a computer, information is stored in the memory which is addressed by its location. Any new information in the same location destroys the old information. In contrast, in a neural network new information is added by adjusting the interconnection strength, without destroying the old information. Thus, the information in the brain is adaptable whereas in the computer it is replaceable.

5) **Fault Tolerance**: - Neural networks exhibit fault tolerance. Since the information is distributed in the connections through the network, even if a few connections are snapped or a few neurons are not functioning, the information is still preserved due to the distributed nature of the encoded information. But, computer, on the other hand, inherently not fault tolerance. The information corrupted in the memory can not be retrieved.

### 1.5.2 ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network can be seen as a set of simple units named **neurons**, interconnected to give an oriented architecture. The generated neurons have inputs and outputs, and the system variables are the weights of the connections.
between neurons. The function of a neuron is to gather information via a weighted sum of its inputs. A non-linear transfer or activation function sends the results to the subsequent units until the result reached the output units. Thus, a neural network can be thought as a graphical notation for an equation that states how output values are generated from input values. The main characteristics of such network is their ability to learn to fix their weights in addition to its theoretical and practical advantages:

**Development Speedup:** ANNs speedup the development of new application on the contrary to the analytical model, by relaxing the need of developing efforts and programming skills.

**Good Performance:** ANNs can handle highly non-linear problems, making them vastly superior to classical linear approaches and often making them achieving better performances than achieved through the analytical model in practical applications.

**Online adaptation:** ANNs are able to adapt online, coping with processes fluctuations and improvement their internal knowledge on the field.

**Exploitation speedup:** ANNs, with their repetitive structure, are well suited to software and hardware implementation speedup (Parallelism, VLSI design etc).

The learning algorithms of ANNs are often classified into two categories – *supervised* and *unsupervised*. The objective of these algorithms is to choose the suitable connection weights to perform a specific task.

**Supervised Learning:** Here weights are adjusted on the basis of direct comparison of the output of the network with known correct answers. The training data is a set of network input coupled with a corresponding target outputs.

**Unsupervised Learning:** The network is expected to discover some underlying organization of the training data and to adjust automatically its weights.
ANNs have been successfully applied for the solution of variety of problems. Some of the common application domains are:

**Pattern Recognition:** ANNs have shown remarkable progress in the recognition of visual images, handwritten character, printed character, speech and other pattern recognition based tasks.

**Optimization/ Constraints satisfaction:** This comprises problems that need to satisfy constrains and obtain optimal solutions. Examples of such problems include manufacturing scheduling, finding shortest possible tour giving a set of cities etc. Several problems of this nature arising out of industries and manufacturing fields have found acceptable solutions using ANNs.

**Forecasting and risk management:** ANNs have exhibit the capability to predict situations from past trends. They have, therefore, found ample application in areas such as meteorology, stock market, banking and econometrics with higher success rate.

**Control Systems:** ANNs have gained commercial ground by finding applications in control systems. They are widely used in computer based control systems. Beside, they have also been used for the control of chemical plants, robots etc.

### 1.6 REVIEW OF NEURAL NETWORK APPLICATIONS IN SPEECH RECOGNITION

The neural network approaches for speech recognition can be divided into three categories [95]. Each category reflects very different prospective. In the first category, neural networks are use as static pattern classifier [32,70,75,86]. The second category contains the hybrid approaches in which neural network are combined with other temporal processing techniques such as Dynamic Time Wrapping (DTW) and Hidden Markov Model (HMM) [17,18,111,124]. The third
category of approaches are aimed at developing dynamic neural network models which are capable of accepting sequentially presented spectral features and also capturing the temporal relations of these features [19,42,123,138]. The following sections represent some representative work of these three categories and their relative merits and demerits.

1.7 STATIC PATTERN CLASSIFICATION

1.7.1 MULTI-LAYER PERCEPTRON

Multi-layer perceptron (MLPs) with error-back propagation training have been successfully applied in a variety of pattern recognition problems[20,23,136]. They have good discrimination capability and can generate complex nonlinear decision boundaries. All the properties are very useful for speech recognition. MLP may have any number of hidden layers, although a single hidden layer is sufficient for many applications, and additional hidden layers tend to make training slower, as the terrain in weight space becomes more complicated. MLP can also be architecturally constrained in various ways, for instance, by limiting their connectivity to geometrically local areas, or by limiting the values of the weights or trying different weights together.

The experimental work by Elman and Zipser [32] was one of the earliest attempts of applying multi-layer perceptron to classify per-segmented speech unit. A similar approach was adopted by Lippmann and Gold [70] to build a spoken digit recognizer. Other important applications include vowel recognition [53], stop consonants recognition [75], Phoneme classification [86, 101], etc.
1.7.2 RADIAL BASIS FUNCTION NETWORK

For speech recognition, a radial basis function (RBF) network accepts the same kind of static patterns as the MLP. The network is directly minimize the radial distance between input pattern and specific centroid templates. As a result, rapid training can be achieved and discrimination is also improved.

Pioneering application of RBF network for speech recognition was reported by Niranjan and Fallside[86]. They used it to recognize eight English vowels. In their experiment, input feature contains FFT power spectrum averaged over several frames. Another set of application was carried out by Renals and Rohwer [101] to recognize 20 vowels. The input feature contains LPC Cepstral coefficients derived from manually labelled vowel segments. The training time required for the RBF classifier was found to be two to three orders of magnitude shorter than that for a MLP and similar recognition accuracy was obtained.

1.7.3 LEARNING VECTOR QUANTIZER

Learning Vector Quantizer (LVQ), also referred to as self-organized map (SOM), was first introduced by Kohonen [56]. It is operated in a somewhat similar way as the RBF network but trained by an unsupervised competitive learning algorithm. Kohonen [58] also suggested an initial application of LVQ to phoneme classification. Two later versions of LVQ, referred to as LVQ2 and LVQ3, were developed to improve the classification performance by providing a more accurate approximation of the optimal Bayesian decision boundaries[57].

McDermott and Katagiri [80] proposed an LVQ based shift-tolerant phoneme recognition technique. This approach adopted a moving window scheme where
each phoneme was represented using a set of time ordered spectral vectors. For recognition, these vectors were presented one after another and the resulted neurons activation was accumulated temporally. Finally, the phonemes class with the highest accumulated activation signified the recognition result.

1.8 HYBRID APPROACH

The common essences of various hybrid approaches are:

1) A neural network is trained to perform phonetic classification for each short-time frame of speech and the network output is interpreted either as local distance score or a posteriori probability.

2) Some well established techniques such as DTW or HMM are used to model the temporal relations of individual frames.

In other words, neural networks are employed to do the job that they are found most suitable, while temporal processing aspects is left to others.

1.8.1 NEURAL NETWORK AND DYNAMIC TIME WARPING

In conventional DTW recognition methods, frame level spectral distances are computed by comparing the input vector with a set of pre-stored templates. Such a finite number of templates are often inadequate to reflect the large speech variability that may be encountered. In view of this drawback, neural networks can be used as non-parametric models to replace the explicit templates and generate the required local distance score.

Bourlard and Wellekens [18] combines a three-layer neural network and standard DTW algorithm for the recognition of German digits. The MLP has 26 output neurons, each representing an allophone of German. For each input frame,
each output neuron generated an “all-or-none” type distance score. These distance scores were then used as the basis for time alignment and the recognized digit was the one with the minimum total distance.

Wang [124] designed an integrated speaker-independent recognizer for connected Mandarin digits, which contained three major components: (a) a modified LVQ, (b) the one-stage dynamic programming algorithm and (c) a Hopfield network. The LVQ network was used to generate the \textit{a posteriori probability} for each input vector. The Dynamic Programming algorithm utilized the probabilities of successive frames to produce a sequence of candidate digits. As the final stage of recognition, the Hopfield network, being equipped with a special time-alignment energy function, performing the essential post-processing to correct those unlikely candidates. In this method, an average recognition accuracy of 94.3\% was attained.

1.8.2 NEURAL NETWORK AND HIDDEN MARKOV MODEL

Conventional HMM techniques use the maximum likelihood training criterion. The parameters of different models are optimised independently to fit the statistical distributions of their own pattern classes. As a result, the \textit{a posteriori probability} of the desired class is increased but, of course, that of undesired classes may not necessarily be decreased. This implies a poor discrimination between models.

In a hybrid NN/HMM system, a discriminatively trained neural network is used to estimate the local probability for all HMM states. Each output neuron essentially corresponds to a particular state. At the same time, other standard HMM operations such as forward-backward evaluation procedure and Viterbi decoding algorithm are used as usual.
Singer and Lippmann [111] made a successful attempt to build RBF/HMM hybrid recognizer for the nine highly confused E-set ([b], [c], [d], [e], [g], [p], [t], [v] and [y]) alphabets. In this application, each word was modelled by a left-to-right HMM with 3 states. Therefore, the RBF network had totally 27 output neurons. The hidden neurons were divided into three groups to deal with the original cepstral vector, first order cepstral vector and second order Cepstral vector respectively. Each group contained 33 Gaussian basis functions. After fine tuning, the final recognition rate was 88.5% in the speaker independent test which is higher compared to 84.3% accuracy in standard HMM with 8 states per word.

Bourlard and Morgan [17] described their extensive theoretical and experimental work on the hybrid use of MLP and HMM. To avoid the necessity of manual segmentation of speech data, it was proposed to embed the MLP training in a Viterbi algorithm and iteratively revised an initial segmentation during the training process. For the recognition of 200 German sentences spoken by a single speaker, the hybrid approach attained a recognition rate of 65.3% against the performance of conventional HMM as 56.9%.

1.9 DYNAMIC NEURAL NETWORK

Dynamic neural network models have been developed specially for recognizing dynamic speech. Time delay connections or feedback connections are utilized so that the network response depends not only on the current input frame but also on the past frames. Dynamic models are capable of accepting sequential input patterns directly.
1.9.1 TIME DELAY NEURAL NETWORK

The time delay neural network (TDNN) employs the basic structure of MLP and incorporates fixed delay lines to provide temporal context from lower layers to higher layers [125]. Different layers may have different number of delays, depending on applications. The recognition decision is made according to the neuron activations at the output layer accumulated over a fixed time window. The TDNN can be trained using a slightly modified version of *Error Back-Propagation* algorithm.

TDNN was proposed and first utilized by Waibel [125] to classify three voiced consonants [b], [d] and [g]. The input layer has 16 neurons (correspond to 16 Mel-scale spectral features), each with 2 delays. The two hidden layers contains 8 and 3 neurons with 4 and 8 delay respectively. Three output neurons were used to represent the consonants. In the speaker independent case, an overall recognition accuracy of 98.5% was obtained while the accuracy of an HMM recognizer was only 93.7%. The same type of networks were then scaled up to form a larger phonemic TDNN for the recognition of all Japanese consonants. In a modular network architecture in which separate sub-network were first trained for different consonant groups and combined together eventually, the ultimate recognition rate increased to 95.9%.

For word level recognition, the so-called multi-state TDNN (MS-TDNN) was devised [42, 138]. Instead of directly accumulating the frame level likelihood scores in the output layer, the MS-TDNN included a *state layer* to perform dynamic time wrapping on these scores. Each phoneme was represented by a particular state and may have variable duration. Word recognition was then done by finding the optimal state sequence. In recognizing connected alphabet strings, the MS-TDNN showed an overall accuracy of 93.6% in the case of single speaker.
Strictly speaking, TDNN is not a "real" dynamic network. It still employs a spatial representation of time. Since the number of delays is fixed, the network may have problem in recognizing phoneme which have varied duration.

1.9.2 RECURRENTNEURAL NETWORK

Recurrence neural networks refer to a broad class of neural networks which have delay and feedback connections [72, 99]. With the recurrent connections, the network state depends not only on the present input but also on the previous states. Therefore, RNNs are capable of modelling complex sequential relations. RNNs have been widely applied in time series predictions [25], grammatical inference [137], dynamic system identification and control [49], etc. For speech recognition, however, the use of RNNs are not as popular as TDNN and other NN approaches described above. This is mainly due to two major problems which have not been successfully resolved. Firstly, it is difficult to determine appropriate target output to train an RNN. Such target outputs must be specified at each time instant and are required to have proper temporal meanings. Secondly, to deal with static and dynamic speech features simultaneously, a large network size becomes essential to ensure reasonable convergence. As a result, for large vocabulary applications in particular, the training time is just too long to be acceptable.

Robinson and Fallside [99] described an RNN/HMM hybrid approach for word recognition, in which the RNN was used to estimate phone (the smallest perceptible discrete segment of sound in a continuous stream of speech) probability. Compared with the conventional MLP/HMM methods, the RNN has provided improved contextual modelling at frame level. However, the temporal dependency at phone level were not emphasized sufficiently. For the 1000-word DARPA Resource
Management task, the RNN has 85,400 adjustable weights. It took about one week to train the network on a dedicated T800 transputers which contained a 65-processor array. The best word recognition accuracy of 95.4% was achieved.

Indeed, the dynamic nature of speech signal presents a number of challenging problems which have never appear in other application areas of neural networks. Among the variety of neural network architectures, RNN shows excellent temporal properties that fit exactly to the requirements for modelling speech dynamics. Therefore, it is believe that RNN has good potential to become a powerful speech recognition technique and much work needs to be done in this area.

1.10 ORGANIZATION OF THE THESIS

The thesis is organized into seven chapters. The first chapter of the thesis presents an introduction to speech recognition research and various tools and techniques used for that purpose. The second chapter is devoted to the representation of Assamese and Bodo phonemes in phonological domain. In this chapter the phonemes are classified on the basis of manner of articulation and position of articulation. The third chapter illustrates the acoustical properties of Assamese and Bodo languages. In this chapter each and every phonemes of Assamese and Bodo languages has been represented on the basis of acoustic properties like Formant Frequencies, Spectrogram, Energy, pitch and Duration. The fourth chapter is devoted to the study of the methods and algorithms associated with the extraction of features from speech signal, which works as the front-end of speech recognizer. In the fifth chapter, Multi-layer Perceptron with different number of hidden nodes and layers has been tested as speech recognizer and their performances have been evaluated. In this chapter, two clustering techniques k-mean and self organized map (SOM) have been
used for clustering the feature vector received from feature extractor to reduce their dimensionality and performance have been compared with respect to same MLP configuration. The developments of a speech recognizer that can overcome the problems in the recognition of speech in the present context, due to its time-varying nature have been detailed in the sixth chapter. Recurrence neural network has been used in this chapter to recognize the phonemes of Assamese and Bodo languages and their performance has been evaluated. The seventh chapter presents the concluding remarks based on the present study and various tools and the techniques used during the study with their relative advantages and disadvantages.

Finally, the possible benefits which can be acquired by further extending the present study have been discussed.