CHAPTER-II

LITERATURE REVIEW:

The different approaches to speech recognition by machine as reported by different group of researchers have been discussed in this chapter. It also includes the strengths and weaknesses of each approach with an explanation why the Artificial Neural Network Approach is the most highly relied one.

2.1: SPEECH RECOGNITION APPROACHES:

Speech recognition deals with the recognition of the specific individual speech sounds. The three popular techniques of speech recognition are: Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Artificial Neural Network (ANN).

2.1.1: THE THREE DIFFERENT TYPES OF SPEECH RECOGNITION APPROACHES:

(A) Dynamic Time Warping Approach:

*Dynamic Time Warping* (DTW) approach is one of the oldest and most popular approaches in speech recognition. The simplest way to recognize an isolated word sample is to compare it against a number of stored word templates and determine which word has the “best match” (Figure 2.1). But this goal is
complicated by a number of factors. First, different samples of a given word will have somewhat different durations (temporal variation). This problem can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration.

**Fig: 2.1: Dynamic Time Warping (DTW)**

However, another problem is that the rate of speech may not be constant throughout the word; in other words, the optimal alignment between a template and the speech sample may be nonlinear. DTW is able to achieve promising accuracy of higher than 95% in digit recognition [48]. DTW is only
suitable to be used in the recognition of small vocabulary because it is computational intensive. It is not practicable in real-time system when the vocabulary is large. In order to apply DTW to the word-recognition problem, we need to know the beginning and ending points of the words. In noisy conditions this is not a trivial task.

The advantages of DTW are:

• Easy and efficient hardware implementation.
• The training sequence is simple, since it just involves the feature extraction for words that needed to be recognized.

The disadvantages of DTW are:

• It is not suitable for continuous speech recognition.
• It requires the computation of the beginning and ending points of the word.

(B) Hidden Markov Model (HMM):

Hidden Markov Models (HMM) [49] are essentially statistical models to assign the greatest likelihood or probability to the occurrence of the observed input pattern. It is a doubly stochastic process with hidden underlying process. HMM represents speech by a sequence of states, each representing a piece
of the input signal. The states of the HMM correspond to phones, bi-phones or tri-phones. At each state, there is a probability distribution for each of the possible letters, and a transition probability to the next state. The speech recognition processes then boils down to finding the most probable path. The training procedure for the HMM-based recognizer is more complex than the DTW-based recognizer [49, 50, 51, 52].

The advantages of HMM-based approaches are:

• It is easy to incorporate other information, such as speech and language models.

• Continuous HMM is powerful for continuous speech recognition.

The disadvantages of HMM-based approach are:

• The HMM probability density models (discrete, continuous, and semi-continuous) have suboptimal modeling accuracy. Specifically, discrete density HMMs suffer from quantization errors, while continuous or semi-continuous density HMMs suffer from model mismatch.

• The Maximum Likelihood training criterion leads to poor discrimination between the acoustic models. Discrimination can be
improved using the Maximum Mutual Information training criterion, but this is more complex and difficult to implement properly.

(C) Artificial Neural Network (ANN):

*Artificial Neural Network* (ANN) [53, 54, 55] is the most emerging technology in the speech recognition and classification. An ANN is basically an information-processing system that has certain performance characteristics in common with biological neural networks (Fig: 2.1.2).

![ Biological Neuron](image)

*Fig: 2.1.2: Biological Neuron*
It is a system that processes information in a parallel-distributed manner. Although its major drawback is the long training time, it is still widely applied in the speech recognition system because it offers many advantages such as non-linearity, ability of adaptation or learning, robustness and ability to generalize [56, 57].

Multilayer Perceptron (MLP) [58, 59, 60], is one of the most popular Neural network architectures. A basic architecture of MLP is shown in Figure (2.1.3). It is a supervised learning, which adapts its weights in response to the teacher values of the training patterns. Its backpropagation (BP) learning propagates the errors at the output layer back to the hidden and input layer in order to adjust its weights [61, 62]. It is a universal function approximator, which can solve the problem efficiently. Besides, its fast execution speed makes it practical to be implemented in real-time processing. It is used to perform recognition of speech sounds at phoneme, syllable and even isolated word level [64, 65, 66, 67, 68, 69, 70].
Fig 2.1.3: A basic architecture of Multilayer Perceptron (MLP)
2.1.2: COMPARISON BETWEEN SPEECH RECOGNITION APPROACHES:

*Table (2.1.2)* and *Table (2.1.3)* show the comparison between different speech recognition approaches based on the literature reviews [2, 71].

Table (2.1.2): The comparison between different speech recognition approaches.

<table>
<thead>
<tr>
<th>Speech Recognition Phase</th>
<th>Approach</th>
<th>Relevant Variables/Data Structures</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Sampling</td>
<td>ALL</td>
<td>Analog Speech Signal</td>
<td>Analog Speech Signal</td>
<td>Digital Speech Samples</td>
</tr>
<tr>
<td></td>
<td>DTW</td>
<td>Statistical Features (LPC coefficients)</td>
<td>Digital Speech Samples</td>
<td>Acoustic Sequence Templates</td>
</tr>
<tr>
<td></td>
<td>HMM</td>
<td>Subword Features (phonemes)</td>
<td>Digital Speech Samples</td>
<td>Subword Features (phonemes)</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>Statistical Features (LPC coefficients)</td>
<td>Digital Speech Samples</td>
<td>Statistical Features (LPC coefficients)</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>DTW</td>
<td>Reference Model Database</td>
<td>Acoustic Sequence Templates</td>
<td>Comparison Score</td>
</tr>
<tr>
<td>Training and Testing</td>
<td>HMM</td>
<td>Markov Chain</td>
<td>Subword Features (phonemes)</td>
<td>Comparison Score</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>Neural Network with Weights</td>
<td>Statistical Features (LPC coefficients)</td>
<td>Positive/Negative Output</td>
</tr>
</tbody>
</table>
Table (2.1.3): The performance comparison between different speech recognition approaches.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Performance/ Application</th>
</tr>
</thead>
</table>
| DTW        | • Mostly used for isolated, digit and connected word recognition.  
          | • Small vocabulary size.  
          | • Training is simple    |
| HMM        | • Mostly used for continuous word recognition.  
          | • Large vocabulary size.  
          | • Training is complex.   |
| ANN        | • Mostly used for isolated, connected and continuous word recognition.  
          | • Medium vocabulary size.  
          | • Training is time-consuming. |
Artificial Neural Network (ANN) is extremely a powerful computational device [72, 73, 74]. Their massive parallelism makes them very efficient. They can learn and generalize from training data. They are particularly fault-tolerant. Besides, they are also noise-tolerant. In principle, they can do anything that a symbolic or logic system can do. There are many forms of ANN. Most operate by passing neural activations through a network of connected neurons such as MLP (Multi Layer Perceptron), SOM (Self Organizing Map) and Hopfield network. One of the most powerful features of neural networks is their ability to learn and generalize from a set of training data. They adapt the weights of the connections between neurons so that the final output activations are correct.

The goal of the network is to learn some association between input and output patterns. This learning process is achieved through the modification of the connection weights between units. In statistical terms, this is equivalent to interpreting the value of the connections between units as parameters to be estimated. The model of network specifies the learning algorithm to be used. In the section below we will briefly review the fundamentals of neural networks:
2.2.1: PROCESSING UNITS:

A neural network contains potentially huge number of simple processing units. All these units operate simultaneously, supporting massive parallelism. All computation in the system is performed by these units. At each moment in time, each unit simply computes a scalar function of its local inputs, and broadcasts the result to its neighboring units. A basic neuron processing unit is shown in Figure 2.2.1. The units in a network are typically divided into input units, which receive data from the environment; hidden units, which may internally transform the data representation; and/or output units, which represent decisions or control signals.

Fig: 2.2.1: A basic neuron processing unit.
2.2.2: CONNECTIONS:

The units in a network are organized into a given topology by a set of connections or weights. Weights are usually one-directional (from input units towards output units), but they may be two-directional, especially when there is no distinction between input and output units. Weights can be changed as a result of training, but they tend to be changed slowly, because accumulated knowledge changes slowly. A network can be connected with any kind of topology. Common topologies include unstructured, layered, recurrent, and modular networks Fig: (2.2.2.) Each kind of topology is best suited to a particular type of application.

Fig: 2.2.2: Neural network topologies: (a) Unstructured, (b) Layered, (c) Recurrent and (d) Modular.
2.2.3: COMPUTATION:

Computation always begins with presenting an input pattern to the network. Then, the activations of all of the remaining units are computed, either synchronously or asynchronously. In layered networks, it is called forward propagation, as it progresses from the input layer to the output layer. In feedforward networks, the activations will be stabilized as soon as the computations reach the output layer but in recurrent networks, the activations may never be stabilized.

2.2.4: TRAINING:

Training a network means adapting its connections so that the network exhibits the desired computational behavior for all input patterns. The process usually involves modifying the weights but sometimes it also involves modifying the actual topology of the network. In a sense, weight modification is more general than topology modification. However, topological changes can improve both generalization and the speed of learning. In general, networks are nonlinear and multilayered, and their weights can be trained only by an iterative procedure, such as gradient descent on a global performance measure. This requires multiple passes of training on the entire training set; each pass is called iteration or an epoch. Moreover, the weights must be modified very gently so as not to destroy all the previous learning. A small constant called the learning rate is used to
control the magnitude of weight modifications. Finding a good value for the learning rate is very important. If the value is too small, learning takes forever; but if the value is too large, learning disrupts all the previous knowledge. Unfortunately, there is no analytical method for finding the optimal learning rate. It is usually optimized empirically by trying different values.

2.3: TYPES OF NEURAL NETWORKS:

Now we will give an overview of some different types of networks. This overview will be organized in terms of the learning procedures used by the networks. There are three main classes of learning procedures. Most networks fall into one of these categories, but there are also various networks, such as hybrid networks which straddle these categories.

2.3.1: SUPERVISED LEARNING:

Supervised learning means that a “teacher” provides output targets for each input pattern, and corrects the network’s errors explicitly. This paradigm can be applied to many types of networks, both feed-forward and recurrent in nature.

Perceptrons [75] are the simplest type of feed-forward networks that use supervised learning. A perceptron is comprised of binary threshold units arranged into layers Fig: (2.3.1(a)). 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. 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 tying different weights together.

Multilayer Perceptron (MLP) shown in Figure (2.3.1. (b)) can theoretically learn any function, but they are more complex to be trained. However, if an MLP uses sigmoid function rather than threshold function, then it becomes possible to use partial derivatives and the chain rule to derive the influence of any weight on any output activation, which in turn indicates how to modify that weight in order to reduce the network’s error. This generalization of the Delta Rule is known as backpropagation.

![Perceptrons: (a) Single-layer Perceptron (b) Multilayer Perceptron](image)

**Fig: 2.3.1: Perceptrons:** (a) Single-layer Perceptron  
(b) Multilayer Perceptron
Hopfield (1982) [76] studied neural networks that implement a kind of content-addressable associative memory. He worked with unstructured networks of binary threshold units with symmetric connections, in which activations are updated asynchronously. This type of recurrent network is now called a **Hopfield network**.

### 2.3.2: SEMI-SUPERVISED LEARNING:

In semi-supervised learning, an external teacher does not provide explicit targets for the network's outputs, but only evaluates the network's behavior as "good" or "bad". The nature of their environment may be either static or dynamic, as example, the definition of "good" behavior may be fixed or it may change over time. The problem of semi-supervised learning is reduced to the problem of supervised learning, by setting the training targets to be either the actual outputs or their negations, depending on whether the network's behavior was judged "good" or "bad". The network is then trained using the Delta Rule, where the targets are compared against the network's mean outputs, and error is backpropagated through the network if necessary [77]
2.3.3: UNSUPERVISED LEARNING:

In unsupervised learning, there is no teacher, and a network must detect regularities and similarities in the input data by itself. Such self-organizing networks can be used for *compressing, clustering, quantizing, classifying, or mapping* input data. This type of network is often called an *encoder*, especially when the inputs or outputs are binary vectors. We also say that this network performs dimensionality reduction.

There is one type of the unsupervised networks which is based on competitive learning, in which one output unit is considered the "winner"; these are known as *winner-take-all networks*. The winning unit may be found by lateral inhibitory connections on the output units. Competitive learning is useful for clustering the data, in order to classify or quantize input patterns [78]. Kohonen [79, 80, 81], developed a competitive learning algorithm which performs feature mapping called *Self-Organizing Map* (SOM). SOM is a neural network that acts like a transformer which maps an m-dimensional input vector into n-dimensional space while locally preserving the topology of the input data. This is the reason that explains why a SOM is called a *feature map*: relevant features are extracted from the input space and presented in the output space in an ordered manner. It is always possible to reverse the mapping and restore the original set of data to the original m-dimensional space with a bounded error. The bound on this error is determined by the architecture of the network and the number of neurons.

39
2.3.4: HYBRID NETWORKS

Some networks combine supervised and unsupervised training in different layers [82, 83]. Most commonly, unsupervised training is applied at the lowest layer in order to cluster the data, and then backpropagation is applied at the higher layer to associate these clusters with the desired output patterns. The attraction of hybrid networks is that they reduce the Multilayer backpropagation algorithm to the single-layer Delta Rule, considerably reducing training time. On the other hand, since such networks are trained in terms of independent modules rather than as an integrated whole, they have somewhat less accuracy than networks trained entirely with backpropagation.

2.4: APPLICATION OF ANN BY DIFFERENT RESEARCHERS:

Many researchers tried to apply neural networks approaches to speech recognition. This is because speech recognition is a pattern recognition task, and neural networks are good in pattern recognition. The earliest attempts involved highly simplified tasks as example, classifying speech segments as voiced/unvoiced, or nasal/fricative/plosive. Success in these experiments encouraged more researchers to move on to phoneme or sub-words classification. The same techniques also achieved some success at the level of word recognition, although it became clear that there were scaling problems when scaling to level of sentences or larger vocabulary size.
Basically, there are two approaches to speech classification using Neural Networks: **static and dynamic**. In static classification, all of the input speech are fed into the neural network at once, and then makes a single decision to classify the speech. By contrast, in dynamic classification, only a small window of the speech are fed into the network, and this window slides over the input speech while the network makes a series of local decisions. These local decisions then have to be integrated into a global decision at the final stage. Static classification works well for phoneme recognition, but it scales poorly to the level of words or sentences. But dynamic classification can scale better than static classification. In the section below we will review some researches in static approach for phoneme/subword classification and word classification.

### 2.4.1: PHONEME/SUBWORD CLASSIFICATION:

Huang and Lippmann [84] have performed a simple experiment to show that neural networks can form complex decision surfaces from speech data. They used a MLP with only 2 inputs, 50 hidden nodes, and 10 outputs, to Peterson and Barney's [85] collection of vowels produced by men, women, & children, using the first two formants of the vowels as the input speech representation. After 50,000 iterations of training, the network produced the decision regions shown in Figure 2.4.1.
These decision regions are nearly optimal, resembling the decision regions that would be drawn by hand, and they yield classification accuracy comparable to that of more conventional algorithms. Figure (2.4.1) Decision regions formed by a 2-layer Perceptron using backpropagation training and vowel formant data. Elman and Zipser [86] trained a network to classify the vowels /a, i, u/ and the consonants /b, d, g/ as they occur in the utterances /ba, bi, bu/, /da, di, du/ and /ga, gi, gu/. Their network input consisted of 16 spectral coefficients over 20 frames and was fed into a hidden layer with between 2 and 6 units, leading to 3 outputs for either vowel or consonant classification. This network achieved an acceptable result with error rates of 0.5% for vowels and 5.0% for consonants. An analysis of the hidden units showed that they tend to be feature detectors,
discriminating between important classes of sounds, such as consonants and vowels. The experimental results demonstrate that backpropagation learning can be used well with complex and natural data.

Among the difficult tasks in classification is the so-called E-set, as example, discriminating between the rhyming English letters “B, C, D, E, G, P, T, V, and Z”. Burr [87] applied a static network to this task, with very good results. His network used an input window of 20 spectral frames, automatically extracted from the whole utterance using energy information. These inputs led directly to 9 outputs representing the E-set letters. The network was trained and tested using 180 tokens from a single speaker. Its recognition accuracy was high which mostly achieved over 99%.

Lee and Ching [88] proposed a design of neural-based speech recognition system for isolated Cantonese syllables. The speech recognition system consists of a tone recognizer and a base syllable recognizer. The tone recognizer adopts the architecture of MLP in which each output neuron represents a particular tone. The syllable recognizer contains a large number of independently trained recurrent networks, each representing a designated Cantonese syllable. A speaker-dependent recognition system has been built with the vocabulary growing from 40 syllables to 200 syllables. In the case of 200 syllables, 3 experiments were conducted on the proposed system and achieved a top-1 (highest result for
experiment 1) accuracy of 81.8% and a top-3 (highest result for experiment 3) accuracy of 95.2%.

2.4.2: WORD CLASSIFICATION:

Peeling and Moore [89] applied MLP to digit recognition with excellent results. They used a static input buffer of 60 frames (1.2 seconds) of spectral coefficients, long enough for the longest spoken word; longer words were padded with zeros and positioned randomly in the 60-frame buffer. Evaluating a variety of MLP topologies, they obtained the best performance with a single hidden layer with 50 units. Comparison is made between proposed MLP and HMM where error rates were 0.25% vs. 0.2% in speaker-dependent experiments, 1.9% vs. 0.6% for multispeaker experiments using a 40-speaker database of digits. In addition, the MLP was five times faster than the HMM system.

Kammerer and Kupper [90] applied a variety of networks to the TI 20- word database, finding that a Single-layer Perceptron (SLP) outperformed both MLP and DTW template-based recognizer in many cases. They used a static input buffer of 16 frames, into which each word was linearly normalized, with 16 2-bit coefficients per frame. Error rates for the SLP vs. DTW were 0.4% vs. 0.7% in speaker-dependent experiments, or 2.7% vs. 2.5% for speaker-independent experiments.
Burr [91] applied MLP to the more difficult task of alphabet recognition. He used a static input buffer of 20 frames, into which each spoken letter was linearly normalized, with 8 spectral coefficients per frame. Training on three sets of the 26 spoken letters and testing on a fourth set, an MLP achieved an error rate of 15% in speaker-dependent experiments, matching the accuracy of a DTW template-based approach.

Kohonen [92, 93] has described a microprocessor-based real-time speech recognition system. It is able to produce orthographic transcriptions for arbitrary words or phrases uttered in Finnish or Japanese. It can also be used as a large-vocabulary isolated word recognizer. The acoustic processor of the system transcribing speech into phonemes is based on neural network principles. The so-called Phonotopic Maps constructed by a self-organizing process are employed. The co-articulation effects in phonetic transcriptions are compensated by means of errors at the acoustic processor output. Without applying any language model, the recognition result is correct up to 92% to 97% referring to individual letters.

Ha-Jin Yu and Yung-Hwan Oh [94] proposed a sub-word based neural network model for continuous speech recognition. The system consists of three modules, and each module is composed of simple neural networks. The speech input is segmented into non-uniform units by the network. Non-uniform unit can model phoneme variations which spread for several phonemes and between words. The second module recognizes segmented units. The unit has stationary and
transition parts, and the network is divided according to the two parts. The last module spots words by modeling temporal representation. The results showed that the system can model such phoneme variations successfully. In this research, the recognizer was built by using simple structures of neural networks. The system consists of three modules. The input speech is segmented by the first module, and is classified by the second module. In this research, a module is added to detect words from the result of sub-word unit recognition. The units are trained by the result of word detection, rather than the result of unit recognition itself.

2.4.3: CLASSIFICATION USING HYBRID NEURAL NETWORK APPROACHES:

Keun-Rong Hsieh and Wen-Tsuen Chen [95] proposed a neural network architecture which combines unsupervised and supervised learning for pattern recognition. The network is a hierarchical self-organization map, which is trained by unsupervised learning first. When the network fails to recognize similar patterns, supervised learning is applied to teach the network to give different scaling factors for different features so as to discriminate similar patterns. Simulation results showed that their proposed model obtained good generalization capability as well as sharp discrimination between similar patterns. Salmela et al. [96] proposed a neural network, which is capable of recognizing isolated spoken numbers speaker independently. The recognition system is hybrid architecture of
SOM and MLP. The SOM maps the feature vectors of a word in a constant
dimension matrix, which is classified by MLP. The decision borders of the SOM
were fine-tuning with Learning Vector Quantization (LVQ) algorithm, with
which the hybrid achieved over 99% recognition out of 1232 test set samples. The
training convergence of the MLP was tested with two different initialization
methods.

Kusumoputro [97] proposed an adaptive recognition system, which is
based on Kohonen Self-Organization Map (KSOM). The goals in their research
on ANN are to improve the recognition capability of the network and at the same
time minimize the time needed for learning the patterns. The goals could be
achieved by combining two types of learning: supervised learning and unsupervised
learning. They developed a new kind of hybrid neural learning system, combining
unsupervised KSOM and supervised back-propagation learning rules. The hybrid
neural system will henceforth be referred to as hybrid adaptive SOM with winning
probability function and supervised BP or KSOM (WPF)-BP. This hybrid neural
system could estimate the cluster distribution of given data, and directed it into
predefined number of cluster neurons through creation and deletion mechanism.
The result of experiment showed that the hybrid neural system of KSOM-BP with
winning probability function has higher recognition rate compared to that of
previous KSOM-BP, even using smaller number of cluster neurons. Tabatabai et al.
[98] proposed a hybrid neural network which consists of a SOM and a Perceptron.
The hybrid is proposed for speaker independent isolated word recognition. The novel idea in their system is the usage of a SOM as the feature extractor which converts phonetic similarities of the speech frames into spatial adjacency in the map. The property simplifies the classification task. The system performance was evaluated for recognition of a limited number of Farsi words (numbers "zero" to "ten"). The overall performance of their hybrid recognizer showed to be 93.82%. The benefits of their system are speed and simplicity.

2.5: SUMMARY:

In this chapter, some popular approaches for speech recognition system have been reviewed, and then comparison between the differences of these approaches is being made. Besides, some fundamentals of neural network are reviewed, based on the topology and type of learning. Lastly, some related researches are included in the last section in order to compare and show the efficiency of different approaches or classifications used and the result obtained.