CHAPTER-I

1.1: INTRODUCTION:

Speech recognition is the process of converting an acoustic signal captured by a microphone or a telephone to a set of words. The recognized words can be applied for many applications such as Commands & Control, Data entry, and Document preparation. Speech recognition systems can be characterized by many parameters such as, Training data, Acoustic models, Lexical models, Language models, Speech representation, Modelling Classifications and Search. The speech recognition system can be classified as: (i) An isolated -word recognition system, which requires that the speaker pause briefly between words, and (ii) Continuous speech recognition. The spontaneous or extemporaneously generated speech is much more difficult to recognize. One popular measure of the difficulty of the task, combining the vocabulary size and the language model-called Perplexity, loosely defined as the geometric mean of the number of words that can flow a word after the language model has been applied. Further, there are some external parameters that can affect speech recognition system performance, including the characteristics of the environmental noise and the type and the placement of the microphone. In fact, the speech recognition is a difficult problem, largely because of the many sources of variability associated with the signal. First, the acoustic realizations of phonemes, the smallest sound units of which words are composed, are highly dependent on the context in which they appear. These
phonetic variabilities are exemplified by the acoustic differences of the phonemes such as /t/ in two, true, and butter. **Second** acoustic variability can result from changes in the environment as well as in the position and characteristics of the transducer. **Third**, within-speaker variability can result from changes in the speaker's physical and emotional state, speaking rate, or voice quality. **Finally**, differences in sociolinguistic background, dialect, and vocal tract size and shape can contribute to **across-speaker variability**.

A typical Speech recognition system attempts to model the sources of variability in several ways. At the level of signal representation, researchers have developed representations that emphasize perceptually important speaker-independent features of the signal, and de-emphasize speaker-dependent characteristics. At the **acoustic phonetic level**, speaker variability is typically modelled using statistical techniques applied to the large amount of data. **Word level variability** can be handled by allowing alternate pronunciations of words in representations known as pronunciation networks. The dominant recognition paradigm in the past fifteen years is the **Hidden Markov Models (HMM)**. It is a **doubly stochastic model**, in which the generation of the underlying phoneme string and the frame-by-frame, surface acoustic realizations are both represented probabilistically as **Markov process**. **Neural Networks (NN)** are other option used to estimate the frame based scores, these scores are then integrated into HMM-based system architectures, in what has come to be known as **hybrid system**.
A different aspect of speech recognitions is to facilitate for people with functional disability or other kinds of handicap [1]. If their daily chores is being controlled by voice, definitely it would be a great help for them. With their voice they could operate the light, fan and other switches make turn off/on or any kind of machine can be operated for their domestic or official purposes. This leads to the discussion about intelligent home where these operations can be made available for the common man as well as for handicapped. Even a doctor can use this technique to help him/her in the operation theatre.

With the information presented so far one question comes naturally; how is speech recognition done? To get knowledge of how speech recognition problems can be approached today, some basics of speech recognition and a review of some research highlights is presented herewith.

1.1.1: BRIEF HISTORY OF SPEECH RESEARCH:

In 1950, the first attempts were made for automatic speech recognition by machine [2]. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [3]. The system relied on measuring spectral resonance during the vowel region of each digit. In an independent afford at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 mono syllable words [4]. The system again relied on spectral measurements (as provided
by an analog filter bank) primarily during vowel regions. In 1959, at University
College in England, Fry and Denes tried to build a phoneme recognizer to
recognize four vowels and nine consonants [5]. They used a spectrum analyser and
a pattern matcher to make the recognition decision. A novel aspect of this research
was the use of statistical information about allowable sequences of phonemes in
English (a rudimentary form of language syntax) to improve overall phoneme
accuracy for words consisting two or more phonemes. In 1959 another attempt was
made by Forgie and Forgie, at the MIT Lincoln Laboratories, constructing
recognizer for ten vowels embedded in a/b/-vowel-/t/ format in a speaker
independent manner [6].

In the 1960’s, Suzuki and Nakata from the Radio Research
Laboratories in Tokyo, developed a hardware vowel recognizer in 1961 [7]. Sakai
and Doshita, from Kyoto University, presented a phoneme recognizer in 1962 [8].
Nagata and Coworkers, from NEC laboratories, presented a digit recognizer in
1963 [9]. Meanwhile, in the late 1960’s, at RCA Laboratories, Martin and his
colleague worked on the non-uniformity of time scales in speech events. Martin
developed a set of elementary time-normalization methods, based on the ability to
reliability of the recognition scores [10]. In the Soviet Union, Vinstyuk proposed
dynamic programming methods for time aligning a pair of speech utterances [11].
A final achievement of the 1960’s was the pioneering research of Reddy in
continuous speech recognition by dynamic tracking of phonemes [12]. This
research spawned the speech recognition program at Carnegie Mellon University, which to this day remains a world leader in continuous speech recognition systems.

In the 1970s, researchers achieved a number of significant milestones, mainly focusing on isolated word recognition. This effort made isolated word or discrete utterance recognition a viable and usable technology. Velichko and Zagoruyko in Russia [13] helped in the use of pattern recognition techniques in speech recognition. Sakoe and Chiba in Japan [14] showed how to apply dynamic programming methods successfully. Itakura's research in the United State [15] showed how to use Linear Predictive Coding (LPC) in speech recognition task. Important were also IBM contributions to the area of large vocabulary recognition [16-18]. Also, researchers at AT & T Bell Labs began a series of experiments aimed at making speech recognition systems that were truly speaker independent [19]. They used a wide range of sophisticated clustering algorithms to determine the number of distinct patterns required to represent all variations of different words across a wide variety of population.

In the 1980s, scientists all over the globe seem to give more emphasis on word recognition. A wide variety of connected word recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) [20], the one pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in England [21], the
level building approach of Myres and Rabiner at Bell Labs [22] and the frame synchronous level building approach of Lee and Rabiner at Bell Labs [23].

Speech recognition research in the 1980s was characterized by a shift in technology from template-based approaches to statistical modelling methods, especially the Hidden Markov Model approach [24, 25]. In the mid 80s, the approach of employing HMMs has now become widely applied, virtually, in every speech recognition laboratory by the speech researchers across the globe. Another idea that appeared in the arena was the use of Neural Nets (NN) in speech recognition problems. Several new ways of implementing such NN-based systems were also proposed [26, 27]. The impetus given by Defence Advanced Research Projects Agency (DARPA) to solve the large vocabulary, continuous speech recognition problem for defence applications was decisive in terms of increasing the research in the area of speech recognition tasks.

In India, speech research with reference to the different Indian languages is carried out at different Laboratories i.e. C-DAC (Pune), C-DAC (Kolkata), IIT (Kanpur), Gauhati University, TIFR IIT (Delhi) and many more. Studies on the acoustical features of some Indian languages like Tamil, Telugu, Malayalam, Bengali and Hindi is carried out by several groups of workers [28-35]. S Saraswathi and T. V. Geetha described the design of a multilingual speech recognizer using a Large Vocabulary Continuous Speech Recognition (LVCSR) database, Global phoneme database for Dravidian Languages namely Tamil,
Malayalam and Telegu in their paper [36] in 2004. S. Mohanty, S. Bhattacharya and A. K. Senapati developed and tested successfully a phoneme recognition system for Oriya Language Phoneme [37]. T. Pruthi et al have described the implementation of Swaranjali, an experimental, Speaker-dependent, real-time, isolated word recognizer for Hindi [38]. P.H.Talukdar et al [39, 40, 41, 42] made studies on the spectral characteristics of Bodo, Rabha and Assamese languages during the last decade.

1.2: SPEECH PRODUCTION MECHANISM:

Human communication is to be seen as a comprehensive diagram of the process from speech production to speech perception between the talker and listener (Fig. 1.0)

![Fig: 1.0: Schematic diagram of the Speech production / perception process](image-url)
As shown in the diagram, the five different elements, ie A. *Speech formulation*, B. *Human vocal mechanism*, C. *Acoustic air*, D. *Perception of the ear*, E. *Speech comprehension*, are explained in the following sections.

The *first* element (*A. Speech formulation*) is associated with the formulation of the speech signal in the talker’s mind. This formulation is used by the human vocal mechanism (*B. Human vocal mechanism*) to produce the actual speech waveform. The waveform is transferred via the air (*C. Acoustic air*) to the listener. During this transfer the acoustic wave can be affected by external sources, for example noise, resulting in a more complex waveform. When the wave reaches the listener’s hearing system (*the ears*) the listener percept the waveform (*D. Perception of the ear*) and the listener’s mind (*E. Speech comprehension*) starts processing this waveform to comprehend its content so the listener understands what the talker is trying to tell him or her.

One issue with speech recognition is to “*simulate*” how the listener process the speech produced by the *talker*. There are several actions taking place in the listeners head and hearing system during the process of speech signals. The perception process can be seen as the inverse of the speech production process. Worth mentioning is that the production and perception is highly a nonlinear process.
1.2.1: SPEECH PRODUCTION:

To be able to understand how the production of speech is performed one need to know how the human’s vocal mechanism is constructed. All human speech sounds begins as pressure generated by the lungs that pushed air through the vocal tract. The vocal tract consists of the pharynx, the mouth or oral cavity and nasal cavity as shown in Figure 1.1. The sound produced depends on the state of the vocal tract as the air is pushed through it. The state of the vocal tract is determined by the position, shape and size of various articulators such as lips, jaw, tongue and velum. The human speech production mechanism involves the respiration of lungs which provides the energy source, the phonation of vocal cords or folds which act as source of sounds, the resonation of vocal tract which resonates the sounds from the vocal folds and the articulation mechanism at the oral cavity which manipulates the sounds from the vocal folds into various distinctive sounds.

The speech sounds can be produced in a relatively open oral cavity or through a constriction in the oral cavity. The speech sounds are produced in a continuous way. As a result, the speech sounds have to be chopped into small units called phones for analysis. Each phone is included in brackets [ ] to indicate that it is a type of sound. The speech sounds are classified into vowels and consonants.
1.3: SPEECH REPRESENTATION:

The speech signal and all its characteristics can be represented in two different domains, the time and the frequency domain. A speech signal is a slowly time varying signal in the sense that, when examined over a short period of time (between 5 and 100 ms), its characteristics are short-time stationary. This is not the case if we look at a speech signal under a longer time perspective.
(approximately time $T > 0.5$ s). In this case the signals characteristics are non-
stationary, meaning that it changes to reflect the different sounds spoken by the
talker.

To be able to use a speech signal and interpret its characteristics in a
proper manner some kind of representation of the speech signal are preferred. The
speech representation can exist in either the time or frequency domain, and in three
different ways [1]. These are a three-state representation, a spectral representation
and the last representation is a parameterization of the spectral activity. These
representations will be discussed in the following sections.

1.3.1: THREE-STATE REPRESENTATION:

The Three-state representation is one way to classify events in
speech. The events of interest for the three-state representation are:

Silence (S) - No speech is produced.

Unvoiced (U) - Vocal cords are not vibrating, resulting in an
aperiodic or random speech waveform

Voiced (V) - Vocal cords are tensed and vibrating periodically, resulting in a
speech waveform that is quasi-periodic. Quasi-periodic means, the speech
waveform can be seen as periodic over a short-time period (5-100 ms) during
which it is stationary.
Fig: 1.2 (a): Time Domain Representation of the Bodo word ‘engkhur’

Fig: 1.2(b): Three-State Representation of the Bodo word ‘engkhur’
The upper plot (a) contains the whole speech sequence and in the middle plot (b) a part of the upper plot (a) is reproduced by zooming in an area of the whole speech sequence. At the bottom of Fig. 1.2, the segmentation into a three-state representation, in relation to the different parts of the middle plot, is given. The segmentation of the speech waveform into well-defined states is not straight forward. But this difficulty is not as a big problem as one can think. However, in speech recognition applications the boundaries between different states are not exactly defined and therefore non-crucial.

As complementary information to this type of representation it might be relevant to mention that these three states can be combined. These combinations result in three other types of excitation signals: mixed, plosive and whisper.

1.3.2: SPECTRAL REPRESENTATION:

Spectral representation of speech intensity over time is very popular, and the most popular one is the sound spectrogram (Figure 1.3). Here the dark blue represents the parts of the speech waveform where no speech is produced and the lighter (red) parts represent intensity if speech is produced. Figure 1.3 (a) shows a spectrogram in the frequency domain and in Figure 1.3 (b) the speech waveform is given in the time domain. For the spectrogram Welch's method is used, which uses averaging modified periodograms [3]. Parameters used in this method are blocksize
Fig: 1.3:  
(a) Spectrogram of the Bodo word ‘Duthang’

(b) Speech amplitude of the Bodo word ‘Duthang’
K = 320, window type Hamming with 62.5% overlap resulting in blocks of 20 ms with a distance of 6.25 ms between blocks.

1.4: PARAMETERIZATION OF THE SPECTRAL ACTIVITY:

When speech is produced in the sense of a time-varying signal, its characteristics can be represented via a parameterization of the spectral activity. This representation is based on the model of speech production.

The human vocal tract can (roughly) be described as a tube excited by air either at the end or at a point along the tube. From acoustic theory it is known that the transfer function of the energy from the excitation source to the output can be described in terms of natural frequencies or resonances of the tube, which is known as formants. Formants represent the frequencies that pass the most acoustic energy from the source to the output. This representation is highly efficient, but is more of theoretical than practical interest. This because it is difficult to estimate the formant frequencies in low-level speech reliably and defining the formants for unvoiced (U) and silent (S) regions.

1.5: PHONEMICS AND PHONETICS:

As discussed earlier in this chapter, the speech production begins in the human’s mind, when he or she forms a thought that is to be produced and transferred to the listener. After having formed the desired thought, he or she
constructs a phrase/sentence by choosing a collection of finite mutually exclusive sounds. The basic theoretical unit for describing how to bring linguistic meaning to the formed speech, in the mind, is called phonemes [47].

Phonemes can be grouped based on the properties of either the time waveform or frequency characteristics and classified in different sounds produced by the human vocal tract. The classification is shown in Figure 1.4. Phonemes can be seen as a way of how to represent the different parts in a speech waveform, produced via the human vocal mechanism and divided into continuant (stationary) or non-continuant parts.

Fig. 1.4: Phoneme Classification
1.5.1: CONTINUANT:

A phoneme is continuant if the speech sound is produced when the vocal tract is in a steady-state. The continuant sounds are vowels, fricatives, affricates and nasals.

(a) Vowels:

Vowels are phonated and are normally among the phonemes with highest amplitude. Vowels can vary widely in duration (typically 40-400 ms) [2] and are spectrally well defined. Vowels are produced by exciting a fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. Variations in the cross-sectional area along the vocal tract determine the formants of the vowels.

Vowels are differentiated by the position of the tongue-hump. The tongue-hump position divides the vowels into three groups: front, middle, back.

(b) Consonants:

All forms of excitation may be involved by consonants. Factors that affect consonants are whether the tone of the vocal cords are present or not, precise dynamic movement of the vocal tract articulators and how the talker articulates. Consonants that are classified as continuant might not require a motion of the vocal tract.
(c) Fricatives:

Fricatives are produced by exciting the vocal tract with steady airstreams that becomes turbulent at some point of constriction. Depending on the form of excitation the fricatives are divided into two groups unvoiced and voiced. Those with simple unvoiced excitation are usually called unvoiced fricatives while those of mixed excitation are called voiced fricatives.

(d) Affricates:

Affricates are formed by transitions from a stop to a fricative. The unvoiced affricate is formed when an unvoiced stop, followed by a transition to the unvoiced fricative are produced. The voiced affricate is formed by producing a voiced stop followed by a vocal tract transition to the voiced fricative.

(e) Nasals:

Nasal consonants are voiced sounds produced by the waveform exciting an open nasal cavity and closed oral cavity. Their waveforms resemble vowels, but are normally weaker in energy due to limited ability of the nasal cavity to radiate sound. When forming a nasal the front of the vocal tract is completely closed either with the lips or the tongue. The velum is opened wide to allow sound propagation through the nasal cavity. Nasal formants have bandwidths that are normally wider than those for vowels. When phonemes precede or follow a nasal
sound, it gives that those phonemes become nasalized. The nasalization produces phonemes with broader bandwidths and is less peaked than those without nasal coupling. This is caused by damping of the formant resonance by the loss of energy through the opening into the nasal cavity [1].

1.5.2: NON-CONTINUANT:

The phoneme is non-continuant when the vocal tract changes its characteristics during the production of speech. For example if the area in the vocal tract changes by opening and closing the mouth or moving our tongue in different states, the phoneme describing the speech produced is non-continuant. The non-continuant sounds are diphthongs, liquids, glides and stops.

(a) Diphthongs:

A diphthong involves a movement from one vowel toward another. When a sound is produced and one wish to determine if a specific part of this sound is a vowel or diphthong it depends on how the sound are produced. If the vocal tract does not maintain a constant shape, or if the sound cannot be sustained without articulatory movement, and both vocal targets are vowels, then the sound is a diphthong.
(b) **Semivowels:**

Semivowels are classified as either *liquids* or *glides*. Liquids have spectral characteristics similar to vowels, but are normally weaker than most vowels due to their more constricted vocal tract. A glide consists of one target position, with associated formant transitions towards and away from the target. Glides can be viewed as transient sounds as they maintain the target position for much less time than vowels.

(c) **Stops:**

Sounds in which the *airstreams enters the oral cavity and is stopped for a brief period are called stops*. Stops are transient, non-continuant sounds that are produced by building up pressure behind a total constriction somewhere along the vocal tract, and suddenly release this pressure. This sudden explosion and aspiration of air characterizes the stop consonants.

**1.6: AIM & OBJECTIVE OF THE PRESENT STUDY:**

Scientists and researchers have been trying to develop a machine for recognition of Speech and has become an important and the most challenging field in context of today’s IT-based global transparency and homogeneous society. By 1990, many researchers had demonstrated the value of *Neural Networks* (NN) for important task like phoneme recognition and spoken digit recognition.
However, it is still unclear whether connectionist techniques would scale up to large speech recognition tasks. There is a large variety in the speech recognition technology and it is important to understand the differences between the technologies. Speech recognition system can be classified according to the type of speech, size of the vocabulary, the basic units and the speaker independence. The position of a speech recognition system in these dimensions determines which algorithm can or has to be used.

Speech recognition has been another proving ground for Neural Networks. Some researchers achieved good results in such basic tasks as voiced/unvoiced discrimination [43], phoneme recognition [44] and spoken digit recognition. However, research in finding a good Neural Network model for robust speech recognition still has a wide potential to be developed.

The aim of the present research work, as embodied in this thesis, is to investigate and formulate a methodology, showing how the Artificial Neural Network (ANN) can be applied in speech recognition. A hybrid model is proposed by combining Self-Organizing Map (SOM) and Multilayer Perceptron (MLP) for Assamese & Bodo speech recognition up to word level.
THE CHIEF OBJECTIVES OF THE PRESENT WORK ARE:

I. Studying the effectiveness of various types of Neural Network (NN) models used in speech recognition.

II. Developing a hybrid model/approach by combining SOM and MLP in speech recognition for Assamese & Bodo language.

III. Making an analysis to determine the optimal values for the parameters (cepstral order, dimension of SOM, hidden-node number, learning rate) in order to obtain the optimal recognition efficiency.

IV. Develop a search technique by using viterbi search for recognition of Assamese and Bodo words.