CHAPTER 1
INTRODUCTION TO SPEECH RESEARCH

1.1 Introduction

Speech is the most natural form of communication for human beings. It is produced as a sequence of sounds. Human Speech is one of the most complex signals to handle. During the last few decades researchers are putting strong efforts in this area with a common consensus that voice recognition may become the next primary user interface. The output of Automatic Speech Recognition (ASR) Technology can be utilized in many applications with a diverse advantage to different professionals like Doctors, Engineers, Lawyers, Teachers, and Students etc. Automatic speech recognition technology permits human speech signals to be used to carry out predefined activities. Once the system detects and recognizes a sound or string of sounds, the recognizer can be programmed to perform according to a predetermined goal.

Many of the developments affecting speech communication have occurred in the last few decades. Since the existence of mechanical speech synthesizers in the 1700s, the basic understanding of the mechanism i.e. how we produce speech took hundreds of years. But the detailed knowledge of audio perception is fairly recent (Bekesy's experiments on the basilar membrane in the 1940s). Modern speech research started around 1930. The invention of the sound spectrograph in 1946 spurred much speech analysis work. It allowed practical displays of the acoustic
output of the vocal tract. More efficient digital speech coding was developed in 1950s in the form of Delta modulation (DM)[12].

During the early 1970's, the Time-adaptive speech coding as well as a large increase in speech recognition work, including the Advanced Research Projects Agency (ARPA)[10], a typical speech understanding project and the use of dynamic programming in matching templates from different speech signals were developed. In the late 1970s, more complex speech systems such as subband and adaptive transform coders[12] appeared. Large-scale integrated circuits made their appearance in the form of on-chip speech synthesizers, and stochastic methods (Hidden Markov Models) became accepted for speech recognition. Some of the major developments of the 1980s included: Single-chip digital signal processors, the use of Vector Quantization for low-rate speech coding, the search for better excitation models for speech synthesis, for example the multipulse excitation model etc, the use of auditory models in speech applications based on hearing experiments using speech like stimuli, and the use of language models to aid speech recognition. During 1990s there has been widespread acceptance of speech coders, synthesizers and recognizers, which have high computational power and are low cost devices. Though the pace of major breakthroughs has slowed in recent years with the introduction of some new techniques such as Artificial Neural Networks (ANN)[13], Fuzzy Logic[6] in 1980s, but still the research continues unabated [1].

Besides speaking face-to-face, speech can also be propagated by other means. In today's world, speech communications has wide application in the area of telephony, recording systems (cassettes, CDs, DVDs, and their players), internet, and many other frontiers of digital world. The design and operation of such systems requires knowledge of
the characteristics of human speech in order to effectively and efficiently convey vocal content. Speaking and listening has come up as the primary means of communication among human although other means of communication techniques, such as sign language, are available. Speaking is achieved through speech. The basic building blocks of the speech of any language are a set of sounds named as phonemes. Without knowing the complexity of speech production, we learn this technique of communication since our early childhood. Even with differences in term of accent, articulation, nasality, roughness, volume, pitch, pronunciation and speed, we are still able to interpret the speech most of the time as long as we are familiar with the language. Since the early childhood our brain learns a spoken language or speech unconsciously. Children learn the basic phonemes within the first year of their birth [2]. Gradually, they start to learn the meaning of words by segmenting sentences and subsequently followed by the development of their vocal tract until they start to understand words and able to pronounce them correctly. Further development continues until the child is able to utter sequences of words to form complete or semi-complete sentences. It is understood that the learning of correct grammar, adaptation to different speakers and environment and even learning of different languages is obvious in the life span of a human being.

With the advent of Speech Technology, the Scientists and Speech researchers are putting strong efforts to produce an efficient speech recognizer so that a natural human-machine interface would be possible and will replace the primitive interfaces, such as keyboard, mouse etc for the computer. This would certainly fuel the human-machine interface and interaction in terms of some valuable applications in our day-to-day life. For examples, language translation machine, smart-home controller and
telephone directory assistance, agricultural assistance improves the quality of living for human. Because of the glamour of designing an intelligent machine that can recognize the spoken language, studies have been done in various fields to achieve this goal [6][7]. To design a successful speech recognition system, expertise and knowledge from a wide range of disciplines are required. Some of the disciplines that have been widely applied to solve the speech recognition problems are: **Signal processing, acoustics, communication theory, computer science, and pattern recognition [70]** etc. Since human brain is efficient in speech recognition, researches have been motivated to build brain-like computational methods.

**1.2 Speech Production Mechanism**

Computers with the ability to understand speech and speak with a human voice are expected to contribute to the development of more natural man-machine interfaces. Computers with this kind of ability are gradually becoming a reality, through the evolution of speech synthesis and speech recognition technologies. However, in order to give them functions that are even closer to those of human beings, we must learn more about the mechanisms by which speech is produced and perceived, and develop speech information processing technologies that make use of these functions. We use speech every day almost unconsciously, but an understanding of the mechanisms on which it is based will help to clarify how the brain processes information and will also lead to the development of more human-like speech devices through the imitation of these functions by computers.
1.2.1 Psychological aspects of Speech Production:

Speech is produced as a sequence of sounds. The articulators such as jaw, tongue, velum, lips, mouth and their shapes, sizes and positions changes over time to produce sounds. Based on the Psychological aspects we can divide the speech production process in three different stages. They are

1. Conceptualization.
2. Formulation.
3. Articulation.

Speech actually starts from our brain as a thought process. It can be considered as a pre-verbal message. This process is known as conceptualization. Second stage of speech production process is speech formulation. In formulation stage our thought (pre-verbal message) is converted into linguistic form. This is known as speech formulation. This stage is again divided into two stages. They are:-

a) Lexicalisation:- Here, our thought will be converted to appropriate words.
b) Syntactic Planning:- Here, the appropriate words will be arranged in the right way (in a syntactically correct way). Articulation is the last stage of speech production. Here, the sound will be produced to convey message.

1.2.2 Biological aspects of Speech Production:

Speech can be defined as waves of air pressure created by airflow pressed out of the lungs and going out through the mouth and nasal cavities. The air passes through the vocal folds (chords) via the path from the lungs through the vocal tract, vibrating them at different frequencies. A simplified diagram of human vocal system is given in Figure (1.1). Some of the main articulators and their functions are also explained.

![Simplified diagram of Human Vocal System](image)

Figure 1.1: Simplified diagram of Human Vocal System
1.2.2.1 The Vocal Tract

The air will undergo further changes as it makes its way upwards towards the mouth, when passed through the larynx. The air that passages above the larynx of a human is called vocal tract. The human vocal tract is classified into two. One is called the oral tract (i.e. pharynx & mouth) and the other is called nasal tract. According to science the average length of the vocal tract of a human being from the larynx to the lips is almost about 17 cm. The pharynx, mouth and nose in the upper cavities of a human are called resonating cavities. The parts of human that produced sound in the vocal tract are called articulators and latter it is subdivided into active (e.g. tongue) and passive (e.g. hard palate).

1.2.2.2 The Glottis

The Epiglottal fold also called Glottis is a leaf-like cartilage. The glottis is attached to the anterior part of human thyroid cartilage and to the root of the tongue of us. The basic function of glottis to cover the entrance to the larynx during swallowing and by covering it prevents food from entering the trachea. Since the glottis is connected to the root of the tongue and hence the whole can be drawn back and down towards the wall of the pharynx to produced special sound.

1.2.2.3 The Pharynx

The pharynx of a human being is approximately 8-12 cm in length of a funnel-shaped muscle. When swallowing a muscle that keeps the passage between the pharynx and the stomach in closed. It always stretches from the larynx and the passage between the stomach & the pharynx.
1.2.2.4. The Velum

The velum of human being is consisted of tissue, blood vessels, nerves, glands and a thin sheet of muscle fibers. When the velum is raised, it presses against the posterior wall of the pharynx, and prevents air from going through the nose which is known as velic closure and that occurs in the production of oral sounds. Again when the velum is lowered, in that case air passes through both the nose and mouth. At some point of our oral cavity when a lowered velum is combined with an obstruction a nasal sound is produced. The basic function of Velum is to separate the nasal cavity from the oral cavity[15]. To produced nasalized speech sound Speakers should habitually leave the velum too much down. It is generally called a nasal ‘twang’. There will be an overall nasalization of the vowels and the failure to pronounce, for instance,/b/, /g/, /d/.

1.2.2.5. The Tongue

The tongue of human is a mobile articulator which is capable of movements of up to 9 times per second. It can take infinite numbers of movement in both vertically and laterally. The versatility can be felt during eating. The main function of tongue is to move the food around in the mouth and pharynx during chewing and swallowing. In the human body tongue is the main principal part to produced vowel sounds.

1.2.2.6. The Nose

The nasal cavity can be divided into two parts – i.e. the nostrils - by a central bone which is generally known as the septum. The roof of the nasal cavity is very narrow, whereas the floor is smooth and relatively wide. The side walls of our nasal cavity are irregular. At the back, the nasal cavity leads into the nasopharynx. The main principal functions of
the nose are the heating of the air during respiration and humidification. The nose or the nasal cavity also acts as a filter.

1.2.2.7 The Teeth and Lips

The teeth is the main part of human to produce sound. The upper teeth are generally used to pronounce the consonants, for example the initial sounds in the English words this (/DIz/) and thing (/TIN/). The lips – which are consisted of glands, tissues and muscles, are used in the formation of both vowels and consonants. For vowels, we have to know whether they are rounded e.g. /u:/ in doom or spread e.g. /i:/ in heed. This two can be pressed together to produce bilabial sounds e.g. /p/, /b/. or when the lower lip articulate with the upper teeth, then also it produce in labiodental sounds e.g. /f/, /v/.

1.2.2.8 The Larynx

It is the principal parts that generate sound where volume and pitch are manipulated. To produce the loudness the lungs expiration strength has a big role. The larynx is manipulated in such a way to generate a source sound with a particular fundamental frequency, or pitch. Based on the position of the tongue, lips, mouth and pharynx the source sound is altered as it travels through the vocal tract.

1.3 Problem Statement

The work embodied in this thesis mainly focussed on the “Analysis and synthesis of Bodo and Assamese Phoneme with respect to their diversity”. The first and foremost task of interest is the speech recognition. The goal of this task is to recognize the spoken language present in the given speech wave or acoustic signal. This task becomes crucial in multi-lingual society like Assam, a north-eastern state of India where Assamese and Bodo are two major link languages. The speech signals comprising
Assamese and Bodo Phonemes are analysed using LPC, MFCC formant frequency techniques and then synthesized using HMM[120] technique. The synthesized speech sound of Assamese and Bodo phonemes are tested in machine by a self developed algorithm.

1.4 Man-Machine Communication and Speech Input

The field of Speech Processing Technology can be broadly divided into the following five categories:

1.4.1 Speech Coding: It is the application of data compression of digital audio signals containing speech. Speech coding uses speech-specific parameter estimation using audio signal processing techniques to model the speech signal. It is then combined with generic data compression algorithms to represent the resulting modeled parameters in a compact bit stream.

1.4.2 Speech Synthesis: It is the process to produce computer-generated simulation of a human speech or voice. Mainly it is used to convert the written information into aural information. For example, a blind person automatically read the content displayed on the screen. The system which is used for this work is called a speech synthesizer which can be implemented in software or hardware products. A text-to-speech (TTS)[40] system converts our language text into speech i.e the given input text is converted to speech sound. By concatenating pieces of recorded speech that are stored in a database, we can synthesize speech. A system that stores phones or diaphones provides the largest output range, but it may lack clarity. A synthesizer can also incorporate a model of the vocal tract and other human voice characteristics to create a fully "synthetic" output voice. The quality of a synthesizer is
determined by clarity or closeness between its ability to be understood and the human voice.

A text-to-speech (TTS) system is generally consists of two parts: one is called **front-end** and the other is called **back-end**. The main function of front-end can be divided into two i.e. First, it translates the raw text consists symbols (say numbers and abbreviations) into equivalent written-out words and the process is called **text normalization**[12] and then it assigns phonetic transcriptions to each word and divide and then it marks the text into prosodic units, say phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme. The Phonetic transcriptions and the prosody information together make up the symbolic linguistic representation that is output by the front-end. The second one is called back-end—generally referred to as the synthesizer—then converts the symbolic linguistic representation into sound.
1.4.2.1 Types of TTS Systems: Typical TTS system can be categorized into the following types depending on its features and applicability.

- **Limited domain**
  - Voice built specifically for an application
    - Limited set of words and sentences
    - Weather forecasts
    - Air/Rail Travel information systems
    - Agriculture information systems etc.

- **Unrestricted**
  - A generic voice capable to reng anything!
    - News Reng
    - Story-telling
    - Desktop assistant etc

1.4.2.2 The different steps followed while synthesizing speech are as follows:

- Record a set of phones say (/a/, /a:/, /i/, /ii/, /k/, /kh/)
- Given a text, for each word obtain the sequence of phones to be concatenated
  - For example: amma → /a/ /m/ /m/ /a/

a) **For Text Input**
- The input text
  - Raw text
  - Formatted text (MS Word, PDF/PS, MS PPT)
  - Tagged text (XML like tags as markup for synthesis)
  - Encoded text
    - Multilingual text in Unicode, Fonts etc, etc.....
b) Conversion from different formats (pdf/ps/doc) to a generic tagged format or raw text

c) Handle Multilingual Text in Unicode

Unicode is similar to ASCII tables, but they can represent practically any language in the world

1.4.2.3 Speech feature to be incorporated

a) Coarticulation
   (i) Coupling effect, when two sounds are produced together
   (ii) Energy
   (iii) Pitch
   (iv) Duration

b) Text Normalization

   It is the process of transforming text into a single canonical form i.e. Text normalization is a process where text is converted in some way to make it consistent in a way which it might not have been before or repeated.

   The Normalizing text before storing or processing it allows for separation of concerns, since input is guaranteed to be consistent before operations are performed on it. It requires being aware of what type of text is to be normalized and how it is to be processed afterwards; there is no all-purpose normalization procedure. It is frequently used when converting text to speech.

c) Abbreviations and Acronyms

   • Title
     • Dr., MD, Mr., Mrs., St. (Saint), etc.
d) Numbers

- Phone numbers
  - +91-40-23001412, (717)-809-8099

- Dates
  - mm/dd/yy, dd/mm/yy, July 4 05, 12-04-05

- Times
  - 13:00, 1:00 PM, 12:15:35

- Money
  - $20, 300 €

- Account numbers
  - 13 digit, 9 digit numbers

- Ordinal numbers
  - 1st, 2nd, 1000th, ½, ¼, 1/100,

- Cardinal numbers - Amounts, statements
  - 2426
    - two four two six
    - twenty four twenty six
    - two thousand four hundred and twenty six

A typical processing of speech feature is given in the figure (1.3).
1.4.2.4 Architecture of a TTS System

1.4.2.5 Linguistic Analysis: The linguistic analysis or content of the speech is analyzed with the aid of the following

(i) Part of Speech (POS) Tagging: To identify

- Proper noun/verb/adjective etc
• A mapping table: word → pos_tag

(ii) Prosodic Phrase breaks
• POS tags are useful to predict phrase breaks in a sentence so that short pauses can be introduced during synthesis

For example:

1. आं ओराहाम जायो ( I eat rice), Here आं - personal pronoun,
   ओराहाम - noun common, जायो - verb
2. तुम्हः आबू मई ताला राम
   तुम्हः [[prop noun] आबू [conj] मई [prop noun]]
   ताला [prop noun] राम [verb]

1.4.2.6 Wave-Form Generation
• Formant Synthesis
• Concatenative Synthesis
  • Diphone synthesis
  • Unit selection synthesis
• Statistical Parametric Synthesis

1.4.2.7 Pronunciation Dictionary
• If no pronunciation dict. then – Use a set of simple rules or to be applied.

• Example: Indian languages
  • A direct correspondence between what is written and what is spoken

For Example:

Hindi Word: namaskaraa → /n/ /a/ /m/ /a/ /s/ /k/ /aa/
   /r/ $

• Note: last /a/ → $ (null)
• /a/ is a short vowel often referred to as schwa
• Process of mapping /a/ → $ is known as schwa deletion
• Schwa deletion can be captured using a set of simple rules
  • Ex: when /a/ occurs at the end of word map it to $
• Letter to Sound rules can be learnt using statistical models too!!
  • CART, HMM, Neural Networks

1.4.2.8 Choice of Unit

Basically there are three ways of choosing units which are to be used in
the process of Analysis & Synthesis of speech. They are

• **Word as a unit**: Its main features are:
  • A large number of units to store
  • Difficult to ensure coverage of all possible words (proper
    nouns etc).
  • Useful for limited domain
  • Phone as a unit: where no coarticulation present.

• **Diphone as a unit:**
  • Preserves the transition region between two phones and
    thus coarticulation is present.
  • Widely used unit for concatenation.

![Diphone Diagram](image)

Figure (1.5): shows the selection & meaning of diphone in a concatenative
speech sound
1.4.2.9 Building a Diaphone Voice

Typically, a diaphone voice can be built through the following steps:

- Record all possible phone-phone combinations in a language
  - Example, record ka, ku, ki, kii, ...kk, ks, kj...
  - Some combinations may not occur!!
- From each of the phone-phone recording, manually label the diphone boundaries
  - Tools such as Emulabel display the waveform and allows you to label the boundaries
- Pool all the diphones to form a diphone database

A Typical Architecture of Diphone Synthesis

![Architecture of Diphone Synthesis](image)

1.5 Speech Recognition Technique

The goal of most of the speech recognition system is to design a model, which can recognize human speech. However, computer based systems do not yet have the capability and flexibility of understanding speech as human does. Human can recognize the sound of interest from a set of sounds audible concurrently. But in a computer based system, the other sounds which are not of interest will be considered as noise. That is
why, improvement in the computer based systems is necessary to make the system as robust as possible.

In speech recognition systems, phonemes and syllables are the most popular sub-word units used [4] [6]. Automatic speech recognition systems are normally designed for three types of utterances. They are:

a) **Isolated word recognition,**

b) **Connected word recognition,** and

c) **Connected speech recognition** also termed **continuous speech recognition**

The first type of recognition is the easiest one. In this case speaker needs to pause automatically in between words. Connected word recognition is capable of analyzing a string of words spoken together, but at a limited speech rate. Connected speech recognition or continuous speech recognition allows the system to recognize normal conversational speech. In this type of system, the system is needed to be trained. Such systems are known as **Speaker Dependent System**[20]. At the same time there are certain systems which are not required to be trained and are termed as **Speaker Independent System**. The main goal of the ASR is the understanding continuous speech, natural or conversational speech. However, in the case of conversational speech, which is spoken in a natural flow, it is very difficult to recognize the words or phrase as there is a very negligible pause or even no pause between words or phrase. So, to distinguish words or phrases, the recognizer need to apply the concept of “guessing”, where the statistical analysis takes place to produce the most likely words or phrases to produce a correct sentence [7].
1.6 Analysis of speech sounds

1.6.1 Spectrogram:

In sound the spectrogram or the sonogram is a visual presentation of spectrum of frequencies, which vary with time or variables. The sonogram also called as spectral water falls or voice grams. The spectrogram are mainly used in the propose of analysis the different calls of animals and to identify the spoken words phonetically they are also used in the progress of music. Spectrograms are generally produced in two ways approximated as a filter bank which output as a series of band pass filter or computed from the time signal using FFT. These two techniques produced two different time frequency distribution. The band pass filters used analog processing to divide the input signal into frequency banks. The typical spectrogram of short section of a vowel is as shown in figure 1.7

![Spectrogram of short section of vowel](image)

Figure (1.7): Spectrogram of short section of vowel

1.6.2 Hidden Markov Model to Phoneme reorganization:

The most famous statistical model of a sequence of feature vector observation is the Hidden Markov Model (HMM). To build HMM recognizer we must decide what sequence will corresponds to what models. The probability metric due to each model is calculated to recognize an utterance and the model is chosen which is best fit to the utterance. But this is inflexible because it requires the new models to be
trained in every time when new words are to be added to the recognizer. To get a general approach a sub word unit called phoneme is used in this model.

Each phoneme model will be made up of a number of states; the number of states per model is another design decision which needs to be made by the system designer. Each state in the model corresponds to some part of the input speech signal. When phoneme based HMMs are being used, they must be concatenated to construct word or phrase HMMs. For example, an HMM for 'cat' can be constructed from the phoneme HMMs for /k/ /a/ and /t/. If each phoneme, HMM has three states the 'cat' HMM will have nine states.

1.6.3 Phoneme Recognition with an Artificial Neural Network (ANN):

Using the error back-propagation technique an artificial neural network has been trained to recognize the phonemes. To extract seven quasi-phonetic features from the spectral frames of a Bark-scaled filter bank a coarse feature network should be trained. The coarse features were recognized with 80% - 93% and in manual segmentation the phone recognition rate was 64% and in 82%.

1.7 Sound Features.

1.7.1 The Vowels: Vowels are produced when the airstream is voiced due to the vibration of the vocal cords in the larynx, and then it shaped using the tongue and after that the lips to modify the overall shape of the mouth. For example if we try to pronounce “mur” out loud, we should be able to feel that our tongue changes position in our mouth.

1.7.2 Diphthongs: Diphthongs are described on the basis of the tongue movement from the beginning to an end position since diphthongs
are gliding sounds. For example the phone /au/ is a rising diphthong that started from the position of the vowel /a/ and ends at /u/. In practice the glide is hardly ever long enough for the full second sound to be reached, and in front of Fortis consonants the glide is particularly short. When followed by Lenis consonants, the first element of the diphthong is considerably lengthened.

1.7.3 Nasal Consonants: A nasal consonant which is also called nasal stop or nasal continuant are produced by lowering the velum in the mouth and allowing air to escape freely through the nose. The oral cavity which acts as a resonance chamber for producing the sound. The air does not escape through the mouth as it is blocked by the lips or tongue.

1.7.4 Fricatives: Fricative: The consonant sound are produced by bringing the mouth into position to block the passage of the airstream, so that air moving through the mouth generates audible friction. Fricatives can be produced with the same positions of the vocal organs as stops; labiodental, alveolar, dental, bilabial, palatal, uvular and velar consonants.

1.7.5 Stops: In phonetics, a stop, also known as a plosive, is an oral occlusive, a consonant in which the vocal tract is blocked so that all airflow ceases. The occlusion may be done with the tongue lips or glottis. Stops contrast with nasals, where the vocal tract is blocked but airflow continues through the nose and with fricatives, where partial occlusion impedes but does not block airflow in the vocal tract.
1.8 Mathematical concept used in Speech Processing

1.8.1 Simple signals: A signal as referred to as a function that conveys information about the behavior or attributes of some phenomenon. It is a function of time that specifies a unique value or amplitude for every instant of time. Such functions are described by a correspondence or mapping, relating one set of numbers (the independent variable, e.g., time) to another set (the dependent variable, e.g., amplitude). Continuous or analog signals map a real-valued continuous-time domain into a range of real or complex amplitudes.

Figure 1.8: For Bodo vowel /a/

Figure 1.9: Assamese vowel /a/
1.8.2 Filters and convolutions: Speech processing often employs filters to modify (e.g., attenuate or amplify), as a function of frequency and energy in a signal. In their output or response, lowpass filters preserve only low-frequency components of an input signal, while highpass filters reduce energy at low frequencies. Bandpass and bandstop filters preserve or eliminate, respectively, signal components in specific ranges of frequencies.

1. Low-pass filter removes the frequencies which are greater than some fixed value.

2. High—It removes some frequencies which are lower than some fixed value.

3. Stop—The stop band filter removes the frequencies in range of values.

The frequencies values are fixed in between 0.0 and 1.0. Here 1.0 represents half the sampling frequency: $f/2$ and hence a gives frequency is expressed in terms of this fixed values say for example $1000KHz = 1000/(f/2)$. The filter mentioned here are expressed in two vectors ($[b, a] = [b = \text{numerator, } a = \text{denominator}]$) In 1-D audio wave form the MatLab program provides a function filfilt that takes as arguments like $[b, a]$ and value which denote the order of the filter with the help of frequent function, filter’s frequency response are plotted. Magnitude values at 0 db are unaffected and below 0 db are suppressed.

1.8.3. Sampling:

Signals like speech, music, sensor outputs, etc., are broadly classified as continuous-time signals (CT) or discrete-time (DT) signal depending on whether the times for which the signal is defined are continuous or
discrete. Correspondingly, a CT waveform is referred to as \( s(t) \) or \( x(t) \), etc., where \( t \) is a (continuous) real number, and a DT waveform is denoted as \( s[n] \) or \( x[n] \), etc., where \( n \) is a (discrete) integer, \( n=1,2,3,... \).

Signals like speech, music, sensor outputs, etc., are broadly classified as continuous-time (CT) or discrete-time (DT), depending on whether the times for which the signal is defined are continuous or discrete.

For example, a typical CT sine waveform with amplitude 1 and frequency \( f_0 \) in Hz can be written as \( x(t) = \sin(2\pi f_0 t) \) to obtain a DT sine waveform with the same parameters, \( x(t) \) can be sampled at times \( t = nT_s = n/F_s \), where \( F_s = 1/T_s \) is the sampling frequency in Hz and \( T_s \) is Sampling Period. This yields \( x(nT_s) = x_n = \sin(2\pi f_0 nT_s) = \sin(2\pi f_0 n/F_s) = \sin(2\pi \phi_0 n) \), where \( \phi_0 = f_0/F_s \) is a normalized (with respect to \( F_s \)) frequency. The signals \( x(t) \) and \( x_n \) are shown in the following graphs for \( f_0 \) and \( F_s = 200 \) Hz.

The analog frequencies are reconstructed at lower frequencies and the alias frequencies are calculated as \( f_s - f_0 = 200 - 1.3 \ast 200 = -60 \) Hz which is verified from the plotting. The negative sign indicates the \( 180^0 \) phase shift. The sampling at Nyquist rate that results in \( \sin(n\pi) \), which are identically zero. That means that we are sampling at zero crossing points and the signal components are fully issued which can be avoided by adding a small phase shift to sinusoid. In case of cosine wave forms (except \( \cos(90n) \)), this type of problem has not arrived.
1.9 Frequency Analysis

An audio or audible frequency is classified as periodic vibration, the frequency which is audible to the normal human. The range of normal frequency for human is 20 – 2000 Hz. Below 20 Hz feel rather than heard and above 20000 Hz can sometimes by some people. High frequencies are the first to be affected by hearing loss due to age and/or prolonged exposure to very loud noises.

Table 1.1 Frequencies and descriptions

<table>
<thead>
<tr>
<th>Frequency (Hz)</th>
<th>Octave</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 to 32</td>
<td>1st</td>
<td>The human threshold of feeling, and the lowest pedal notes of a pipe organ.</td>
</tr>
<tr>
<td>32 to 512</td>
<td>2nd to 5th</td>
<td>Rhythm frequencies, where the lower and upper bass notes lie.</td>
</tr>
<tr>
<td>1024 to 2048</td>
<td>6th to 7th</td>
<td>Defines human speech intelligibility, gives a horn-like or tinny quality to sound.</td>
</tr>
<tr>
<td>4096 to 8192</td>
<td>8th to 9th</td>
<td>Gives presence to speech, where labial and fricative sounds lie.</td>
</tr>
<tr>
<td>16384 to 32768</td>
<td>10th</td>
<td>Brilliance, the sounds of bells and the ringing of cymbals. In speech, the sound of the letter &quot;S&quot; (8000–11000 Hz)</td>
</tr>
</tbody>
</table>

1.9.1 Fourier Transform: Discrete Fourier Transform: In time domain speech is very difficult to analysis, so it must be converted to frequency domain and for that Fourier transform is used to convert the confirm time domain signal to frequency domain signal which consists of different spectral magnitude information say pitch formants etc. Fast Fourier Transform is the foundation for the signal processing tool box which is used to compute the DFT in reduced execution time. The
function `fft()` and `ifft()` are used in MatLab to calculate the DFT and its inverse. The following two functions implement the relationships.

\[
X(k + 1) = \sum_{n=0}^{N-1} x(n + 1)W_N^n \\
x(n + 1) = \frac{1}{N} \sum_{k=0}^{N-1} X(k + 1)W_N^{-kn}
\]  

Where \( W_N = e^{-j\frac{2\pi}{N}} \)

The figure below shows the waveform representation of a sampled signal in time domain and Fourier transform of the same signal in frequency domain.

![Waveform of an i/p signal](image)

**Figure 1.10 Wave form of an i/p signal**

All the Bodo and Assamese vowels and words used were transformed in the same way for processing.

### 1.9.2 Correlation

A typical plot of autocorrelation sequence of a sine wave with frequency 1 Hz and sampling frequency of 16000 Hz is shown below:
The Matlab function `xcorr` is used to estimate the autocorrelation sequence, it gives double number of samples with the signal \( x(i) \). An important point to remember when using the function `xcorr` is that the origin is in the middle of the figure (Here it is at lag=1024).

## 1.10 Applications

Scientists and researchers however have identified a number of situations in which spoken communications with machines would be advantageous.

Some of the common situations are - when the user’s hands or eyes are busy, a limited keyboard and/or screen is available, disabled users, when any individual is not acquainted with the computer system or when the natural language is preferred [11]. Spoken interaction with machines is a situation in which a user’s hands’ and/or eyes are busy performing another task. When users are able to use speech to communicate with a machine, they are free to pay attention to their other important tasks. It has been noticed that a highly accurate speech recognition systems where speech input can be given, leads to higher task productivity and accuracy [11]. Some of the important application areas are:
1.10.1 Application for the Blind:
As the blind cannot communicate with a computer system, like a normal man does, speech synthesis application seems to be very useful for them. Earlier, before the development of synthetic speech system, to help the blind people, the content of the book were recorded in audio form, which can be listen by the blind. When the output from a speech synthesizer is listened for the first time, it may sound intelligible and pleasant.

1.10.2 Applications for the Deafened and Vocally Handicapped:
People who are vocally handicapped and deafened, they face the difficulties of hearing and speaking. So synthesized speech gives them the opportunities to communicate with other people those who are unable to understand the sign language and hence visual information is the main key point for the dumb and deaf persons. Through the key board it becomes much slower than the normal speaking. So by using some predictive input system which always display the most frequent word for any word fragment and by hitting the special key user to accept the prediction[12].

1.10.3 Educational Applications:
Synthesized speech can be used in the many application of Education through this we can programmed for some important works or tasks say pronunciation and spelling teaching. Generally, for the people which are impaired to read this synthesizer help them when they feel themselves very embarrassing to be helped by a teacher. In case of proof range the synthesizer has a great role by connected in the word processor. Speech synthesizer can also help in misspelling.

1.10.4 Applications for Telecommunications and Multimedia:
In the area of multimedia speech synthesizer has a great role. The synthesized speech has been using in different kinds of telephone enquiry
system. But for the case of E-mail which are not possible to read, with synthetic speech message may be listened to normal telephone line.

Interactive Voice Response (IVR) is an automatic speech recognition system is also needed to get the total interactive application of multimedia. Electronic mail has become very usual in last few years. However, it is sometimes impossible to read those E-mail messages. There may be no proper computer available or some security problems exist. With synthetic speech e-mail messages may be listened to via normal telephone line. Synthesized speech may also be used to speak out short text messages (sms) in mobile phones.

1.10.5 Other Applications and Future Directions:

In all kinds of man-machine interactions speech synthesis is used. For example, to give more accurate information of the current situation in warning and alarm systems synthesized speech may be used. Speech synthesizer may also be used to receive some desktop messages from a computer, such as received e-mail or printer activity.

In the future, if speech recognition techniques reach sufficient level, synthesized speech may also be used in language interpreters or several other communication systems, such as videophones, videoconferencing, or talking mobile phones. With talking mobile phones it is possible to increase the usability considerably for example with visually impaired users or in situations where it is difficult or even dangerous to try to reach the visual information. It is obvious that it is less dangerous to listen than to read the output from mobile phone for example when driving a car. The application field for speech synthesis is growing wider all the time which may come also more funds into speech research and development areas.