CHAPTER 1

INTRODUCTION

1.1 Introduction

Among humans, speech has always been the fundamental mode of communication and in case of human-machine interaction; spoken language has been accepted as the natural method. Our ability to communicate with machines and computers is far more difficult and slower in magnitude when carried out through keyboards, mouse and other devices. Thus speech input is an essential component in order to make this communication more user-friendly. Besides, a great deal of information is expressed by humans by means of natural speech. Thus even an illiterate or a person with little knowledge about computers can operate computers using speech commands.

Many physically disabled people in case they are unable to type in keyboard or click in mouse with their hands may use computers with the help of speech input. People who have sufficient expertise with computers can utilize the speech input ability of a computer to significantly speedup documentation, writing e-mail sending and other significant operations using computers. Also there are many advantages to this mode of operation such as it can be used for many situations when the hands are already being used for other important operations such as driving a car. Two good examples for speech based system are speech enabled dialing and Global Positioning System.
Another exciting application called Automatic Spoken Language Translation had emerged with the development of Machine Translation techniques. This allows people from different countries all over the world to communicate freely without any professional translator via speech.

Development of a speech recognition component is one of the fundamental challenges involved in developing a spoken language system. For approximately three decades, research in the field of speech recognition has been ongoing. Much progress has been made during this time span. Started with very small vocabulary, speaker dependent, isolated word recognition systems, the technology today has been moved to large vocabulary systems, capable of recognizing from 50,000 up to 150,000 words.

In-order to work out of the box for any speaker, the systems now being used is speaker independent. In some the system its even speaker adaptive, learning the peculiarities of a person's speech over time. With the introduction of systems by IBM and Dragon, Isolated speech has long yielded to continuous speech in the research environment, and more recently, in the commercial marketplace as well a dramatical reduction in error rates have been observed.

A sum total of about 74 million people in the world are Tamil. Of that, about 63,000,000 are in India; while about 3,600,000 live in Sri Lanka; about 1,500,000 in Malaysia and about 250,000 in Singapore. The remaining Tamil people are scattered over different places around the world. There is a great potential for the application of the speech technology in the context of Tamil speech.
Since, Tamil being the language spoken only among Tamil people, only limited research has been carried out in Tamil speech recognition in comparison to other Indian languages. Tamil language has a sum total of 247 letters, but most of them are derived from the 18 consonants and 12 vowels. The remaining 216 letters are constructed by combining the sounds of a consonant and vowel with each letter having a unique sound[6].

Figure 1.1 Tamil Consonants

Figure 1.2 Tamil Vowels

Any speech may involve the numerals and thus we need to consider the Tamil numerals for the Tamil speech recognition process in addition to the Tamil words. Thus the data base of numerals is a large one as that of the English numerals.
The main objective of this research is to use Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) for developing an intelligent system for Speech recognition in Tamil Words and Numerals.

1.2 Problem Statement

The Tamil speech recognizers currently available are limited by the following constraints:

- Lack of speaker-independency
- Limitation to isolated word recognition
- Limitation to small vocabulary tasks
- Increased training and computational time
**Lack of speaker-independency:** Although researches on speaker-independent speech recognition in Tamil speech domain were conducted, there is still room for improvement on speaker-independent recognition accuracy. It is desirable to use a large number of speech parameters (or features) in speaker-independent recognition. Thus, a technique that can account for modeling multiple parameters is needed. Automatic and efficient algorithms must exist for training and recognition of the model due to the complexities introduced by freeing these constraints and the greater amount of training data available for speaker-independent recognition.

Mel-Frequency Cepstral Coefficients (MFCC) is a powerful technique capable of robust modeling of speech. Hidden Markov Model (HMM) is the currently used acoustic model in Tamil speech recognizers. Though HMM works considerably well in speaker-dependent tasks, the degradation of accuracy in speaker-independent tasks becomes apparent. Besides, HMM causes accuracy degradation due to quantization error. Thus in-order to counter this limitation a more efficient modeling algorithm must be adapted.

**Limitation to isolated word recognition.** Being constrained to isolated word recognition tasks, the existing Tamil speech recognition system mainly focuses on isolated word, digit recognition. Only a limited research on Tamil continuous speech recognition system is present. Thus ways to develop Tamil continuous speech recognition system needs to be investigated in-order to release this constraint.

**Limitation to small vocabulary tasks:** Being limited to small vocabulary recognition task, the current Tamil speech recognizers can recognize, for instance, may be 10 Tamil digits. Word model is used as acoustic model. The researches have been limited in case of
large or medium vocabulary tasks where the sub-word unit modeling is required. Recognition using sub-word units such as phoneme, diphone and triphone lacks research and study. This constraint the recognition tasks to small vocabulary. Applying sub-words as modeling unit is needed in order to establish Tamil speech recognition system for large and medium vocabulary.

**Increased training and computation time:** If there is a case of large database of words present then the methods such as ANN, HMM and BPN will require more number of training so as to recognize the word properly. Thus this situation will lead to wastage of time and will also increase the total computational time of the system. Thus the speech recognition system developed should be able to come up with recognized word within a short span of time.

1.3 Speech signal

Speech is composed of a sequence of sounds and the sounds when shared between humans serves as a symbolic representation of information. In a medium, such as air or water, speech propagates as a longitudinal wave. Depending on the density of the medium the speed of propagation varies. Commonly the amplitude of air pressure variation corresponding to a speech signal is plotted as a function of time. This kind of plot is known as a speech pressure waveform or just a speech waveform[3].
The simplest form of sound that can be described is called a sinusoid. When describing the properties of sound waves, it is a useful starting point since it has such a pure shape.

1.3.1 Properties of Sinusoids

A sinusoid corresponds to a time waveform drawn with a simple smooth up and down movement. There are three kinds of measurement that can be taken from a sinusoid that can completely define the shape of a waveform when they are taken into account together. These measurements are:

- Amplitude: Size of the displacement above and below the mid line of a sinusoid. This signifies the energy in the sound wave and it corresponds to loudness of the sound wave. Amplitude can be measured in many ways; it can be measured in units of pressure as it is related to the size of the pressure variations in the air. More
often we measure amplitude on a logarithmic scale called decibels (dB) which is relative to a standard sound. The dB scale maps directly to the way that humans perceive loudness and hence is found to be very useful form of measure.

- **Frequency**: The number of cycles the sinusoid makes in each second can be termed as the frequency. A cycle consists of an oscillation from the midline to the maximum, down to the minimum and back to the midline. Frequency is measured in Hertz (Hz) or in other words simply cycles per second.

- **Phase**: The position of the starting point of the sinusoid is termed as Phase. The phase of those starting at the maximum is said to be zero while a phase of $\pi$ radians is considered to be for those starting at the minimum. Perceiving the phase of a sinusoid is not possible but the relative changes in phase between two signals can be detected, as the brain determines the location of a sound on the basis of different phases heard at the two ears forms the basis of binaural hearing.

### 1.3.2 Properties of speech signal

Significant energy is present in speech ranging from zero frequency up to around 5 kHz. The speech signal properties changes remarkably as a function of time. The concept of time varying Fourier representation is used to study the spectral properties of speech signal. However, the temporal properties of speech signal such as energy, correlation, zero crossing etc are assumed to remain constant over a short period. That is its characteristics are short-time stationary. Therefore, Speech signal is divided into a number of short duration blocks using hamming window, so that normal Fourier transform can be used. Some of the important properties of speech signals are summarized as follows[4].
• **Bandwidth:** The speech signal has a bandwidth much higher than 4 kHz. In fact, there is still a significant amount of energy in the spectrum for high and even ultrasonic frequencies for the fricatives. However, as we all know from using the (analog) phone, it seems that the speech signal contains all the information necessary to understand a human voice within a bandwidth of 4 kHz.

• **Fundamental Frequency:** The usage of voiced excitation for the speech sound results in a pulse train, which is termed as fundamental frequency. The signal has a fundamental frequency between 80 Hz and 350 Hz and is periodic in nature. While voiced excitation is normally used in case of articulating vowels and some of the consonants, unvoiced excitation (noise) is used for fricatives (e.g. /s/ as in mess, or /f/ as in fish). Usually no fundamental frequency can be detected in such cases. On the other hand, the signal has very high zero crossing rate. Plosives (like /p/ as in put), that uses transient excitation can be best detected in the speech signal by looking for the short silence necessary to build up the air pressure before the plosives are bursted out.

• **Peaks in the Spectrum:** The vocal tract gives a characteristic spectral shape to the speech signal after passing the glottis. If one simplifies the vocal tract to a straight pipe (approximately of 17 cm length), one can see that the pipe shows resonance at certain frequencies. These frequencies are called formant frequencies. The frequency of the formants (especially of the 1st and 2nd formant) change depending on the shape of the vocal tract (the diameter of the pipe varies along the pipe), and
therefore the vowel being articulated is characterized. The peaks in the spectral distribution of energy is given by \((2n - 1) \times 500\) Hz; Where \(n = 1, 2, 3, \ldots\).

- **The Envelope of the Power Spectrum Decreases with Increasing Frequency:**
  The power spectrum of a pulse sequence from the glottis decreases towards higher frequencies at a rate of -12dB per octave. A high-pass characteristic with +6dB per octave is shown by the emission characteristics of the lips show. Thus, an overall decrease of -6dB per octave is the result. Thus with increasing frequency there occurs a decrease in the envelope of the power spectrum of the signal.

### 1.3.3 Speech production

Speech production takes place by means of the biological system of a larynx or sound box, which resides in the throat of the human beings. The fibers provided in the larynx or the sound box are capable of vibrating when the air passes through them. The lung, attached to the larynx, by the windpipe, is the major organ that pumps air through the larynx. The laryngeal fibers or vocal cords are properly capable of vibrating at all frequencies. The range goes from the just audible 20 Hz up to about 11000 Hz in some particular case. The higher frequencies are usually found in children and women while the lower frequencies are found in men.

In terms of Signal Processing, the larynx can be called as the source and vocal tract as the filter. The Source is capable of vibrating at many or may be at all frequencies. The vocal tract which includes the vocal cavity, the mouth cavity & the nasal cavity, filters the sound thus produced to resonate some frequencies & anti resonates others so that a special sound emanates from the lips in the form of speech[1,2].
The length and shape of the vocal tract do not change constantly but remains stable over a scale of millisecond. Hence the characteristics of the filter are stable over the millisecond scale of the millisecond. Thus on small scale a person’s voice has a static nature and on a larger scale has a dynamic nature.

Figure 1.5 Speech production system in Human
Every time when the speech sounds are produced, the air flowing from lungs first passes through the glottis and then the throat and mouth. The speech signal can be excited in three possible ways depending on the speech sound which is being articulated by the speaker.

- **Voiced excitation:** The air pressure forces the glottis, which is initially in the closed condition, to open and close periodically thereby generating a periodic pulse train (triangle–shaped). This fundamental frequency usually lies in the range from 80Hz to 350Hz.

- **Unvoiced excitation:** The glottis is in the open state and the air passes through a narrow passage in the throat or mouth. This results in a turbulence which results in generating a noise signal. The location of the narrowness determines the spectral shape of the noise.

- **Transient excitation:** The air pressure is raised due to a closure in the throat or mouth. The air pressure drops down immediately by suddenly opening the closure.

With some speech sounds these three kinds of excitation occurs in combination. The shape of the vocal tract determines the spectral shape of the speech signal. You change the spectral shape of the speech signal by changing the shape of the pipe (and opening and closing the air flow through your nose) and thus different speech sounds are articulated.
1.3.4 Speech Production Model:

This model works as follows: A pulse generator models voiced excitation by which it generates a pulse train (of triangle–shaped pulses) with its spectrum given by $P(f)$. A white noise generator models the unvoiced excitation with a spectrum, $N(f)$. One can adjust the
signal amplitude of the impulse generator (v) and the noise generator (u) in-order to mix voiced and unvoiced excitation. The output of both generators is then added and fed into the box which models the vocal tract and performs the spectral shaping with the transmission function \( H(f) \). \( R(f) \) models the emission characteristics of the lips.

Hence, the spectrum \( S(f) \) of the speech signal is given as:

\[
S(f) = (v \cdot P(f) + u \cdot N(f)) \cdot H(f) \cdot R(f) = X(f) \cdot H(f) \cdot R(f) \quad (1.1)
\]

We have the following parameters in our speech production model to influence the speech sound:

- the fundamental frequency (determined by \( P(f) \))
- The spectral shaping (determined by \( H(f) \))
- The signal amplitude (depending on \( v \) and \( u \))
- the mixture between voiced and unvoiced excitation (determined by \( v \) and \( u \))

These are the technical parameters describing a speech signal. The parameters given above have to be computed from the time signal to perform speech recognition and then needs to be forwarded to the speech recognizer. The most valuable information for the speech recognizer is contained in the way the spectral shape of the speech signal changes with time.

1.3.4 Speech perception

A schematic view of the human ear showing the three distinct sound processing sections is shown in the figure 1.8, namely: the outer ear consisting of the Pinna, which gathers sound and conducts it to the middle ear through the external canal; the middle ear beginning at the tympanic membrane (eardrum) and including three small bones namely, the Malleus (also called the hammer), the Incus (also called the anvil) and the stapes (also called the
stirrup), which performs a transduction from acoustic waves to mechanical pressure waves; and finally, the inner ear, consisting of the cochlea and the set of neural connections to the auditory nerve, which conducts the neural signals to the brain.

Figure 1.8 Structure of Human Ear

Speech perception model:
A block diagram abstraction of the auditory system is depicted in the figure 1.9. The acoustic wave is transmitted from the outer ear to the inner ear where the sound wave is converted to mechanical vibrations by the ear drum and bone which are ultimately transferred to the basilar membrane located inside the cochlea. The basilar membrane vibrates in a frequency-selective manner along its extent performing a rough (non-uniform) spectral analysis of the sound. A set of inner hair cells are distributed along the basilar
membrane that serves to convert the motion along the basilar membrane to neural activity. This produces an auditory nerve representation in both time and frequency[1].

The processing at higher levels in the brain, shown in Figure 1.9 as a sequence of central processing with multiple representations followed by some type of pattern recognition, is not well understood and we can only postulate the mechanisms used by the human brain to perceive sound or speech. Even so, a wealth of knowledge about how sounds are perceived has been discovered by careful experiments that use tones and noise signals to stimulate the auditory system of human observers in very specific and controlled ways. These experiments have yielded much valuable knowledge about the sensitivity of the human auditory system to acoustic properties such as intensity and frequency.
1.4 Speech Recognition

Speech Recognition is a special case of pattern recognition. The process of converting speech signal to a sequence of words by means of algorithm implemented as a program is called speech recognition. The goal of speech recognition is to provide machine with the ability to "hear," understand," and "act upon" spoken information. Training and Testing are the two phases in Speech recognition. In both phases, the process of extraction of features relevant for classification is common. The parameters of the classification model are estimated using a large number of class examples (Training Data) during the training phase. The feature of test pattern (test speech data) is matched with the trained model of each and every class during the testing or recognition phase[5]. The test pattern is declared to belong to that model which matches the test pattern in the best way.

![Figure 1.10 Speech Recognition phases](image)

1.4.1 Speech recognition concept in Human

The complete process of producing and perceiving speech from the formulation of a message in the speaker's brain, to the speech signal creation, and finally to the
understanding of the message by a listener is depicted in the figure 1.11. The process starts as a message represented somehow in the brain of the speaker in the upper left. The message information can be thought of as having a number of different representations during the process of speech production. For example, initially, the message could be represented as English text.

In order to “speak” the message, the talker implicitly converts the text, that is the one that corresponds to the spoken version of the text, into a symbolic representation of the sequence of sounds. This step, called the language code generator, converts text symbols to the basic sounds of a spoken version of the message by means of phonetic symbols (along with durational information and stress) and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced.

The next step in the speech production process is the conversion to “neuro-muscular controls,” i.e., the set of control signals that enables the neuro-muscular system to move the speech articulators as directed, which includes the tongue, lips, teeth, jaw, and velum, in a manner that is consistent with the desired spoken message sounds and with an emphasis of desired degree. The neuro-muscular controls step results in a set of articulatory motions (continuous control) that moves the vocal tract articulators in a prescribed manner creating the desired sounds.

The final step involved in the Speech Production process is the “vocal tract system” that creates an acoustic waveform, that encodes the information in the desired message into the speech signal by physically creating the necessary sound sources and the appropriate vocal tract shapes over time.
The series of steps from capturing speech at the ear to understanding the message encoded in the speech signal are shown in the speech perception model. The first step includes the conversion of the acoustic waveform to its spectral representation effectively. This is done by the basilar membrane, which acts as a non-uniform spectrum analyzer within the inner ear, which spatially separates the spectral components of the incoming speech signal and thereby analyzing them by what amounts to a non-uniform filter bank. This step is followed by a neural transduction of the spectral features into a set of sound features (or distinctive features as referred to in the field of linguistics) that can be decoded and processed by the brain. The next step in the process is the conversion of the sound features by a language translation process in the human brain, into the set of phonemes, words, and sentences that are associated with the in-coming message.

Finally, the last step involved in the speech perception model is the conversion of the phonemes, words and sentences of the message into an understandable or of similar meaning to that of the basic message in order to be able to take some appropriate action or respond accordingly. In most of the speech perception modules our fundamental understanding of the processes is rudimentary at best, but it is generally agreed that some physical correlate of each of the steps in the speech perception model occur within the human brain. Therefore, the entire model is useful for thinking about the processes that occur.

Between the speech generation and speech perception parts of the model there exists a transmission channel. This transmission channel, in its simplest embodiment, consists of just the acoustic wave connection between a speaker and a listener who are in a common space. Including this transmission channel in our model for the speech chain is essential as
it includes channel distortions and real world noise that make understanding of speech and message more difficult in real communication environments.

1.4.2 Speech recognition by Machine

The main goal of a speech recognition system is to substitute for a human listener, although it is evident that achieving the flexibility offered by human ear and human brain is very difficult for an artificial system. Thus, some constraints are present in speech recognition systems. The process is dealt with in parts to increase the performance of the recognition, and researches are concentrated on those parts.
The working of speech recognition systems is based on the principle of roughly comparing the input data with the prerecorded patterns. The arrangement of these patterns are done in the form of phoneme or word. The pattern to which the input data is most similar is accepted as the symbolic representation of the data by this comparison. To compare raw speech signals directly is a very difficult task.

A preprocessing on the signals is necessary as significant variation of the intensity of speech signals can occur. This preprocessing is followed by Feature Extraction. Short time feature vectors are obtained from the input speech data initially, and then these vectors are compared to the patterns classified prior to comparison.

1.5 Classification of Speech recognition systems

A. Based on types of Speech Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. A single word, a few words, a sentence, or even multiple sentences may be considered as an utterance. The types of speech utterances are:

**Isolated Words:** Usually isolated word recognizers require to have quiet on both sides of the sample window for each utterance. It doesn't mean that it accepts single words alone, but a single utterance at a time is a prerequisite. Though the system proves to be effective
for situations where the user requires to give only one word responses or commands, it is very unnatural for multiple word inputs. Since word boundaries are obvious and the words tend to be clearly pronounced, it is comparatively far more simpler and easier to implement.

**Connected Words:** Though connected word systems (or more correctly 'connected utterances') are somewhat similar to isolated words, they allow separate utterances with minimal pause in between to be 'run-together'.

**Continuous Speech:** Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. In other words, it's basically computer dictation. It includes a great deal of "co articulation", in which adjacent words run together without any apparent division or pauses in between. As continuous speech recognition systems must utilize special methods to determine utterance boundaries, they are comparatively most difficult to create. Confusability between different word sequences grows as vocabulary grows larger.

**Spontaneous Speech:** This kind of speech is natural and not rehearsed. A variety of natural speech features such as words being run together and even slight stutters should be handled by an ASR system with spontaneous speech. Spontaneous (unrehearsed) speech includes false-starts, mispronunciations and non-words.

**B. Based on the type of Speaker Model**

Due to one’s unique physical body and personality all speakers have their special voices. Based on speaker models speech recognition system is broadly classified into two main categories namely speaker dependent and speaker independent.
**Speaker dependent models**: Speaker dependent systems are designed for a specific speaker. For a particular speaker they are generally more accurate, but much less accurate for other speakers. Though these systems are usually cheaper, easier to develop and more accurate they are comparatively not as flexible as speaker adaptive or speaker independent systems.

**Speaker independent models**: Speaker independent systems are designed for supporting a variety of speakers instead of a specific one. It can recognize the speech patterns of a large group of people. But unlike speaker dependent systems this system is more difficult to develop, more expensive and offers comparatively lesser accuracy. However, they are more flexible.

**C. Based on types of Vocabulary**

The complexity, processing requirements and the accuracy of the system is greatly affected by the size of the vocabulary of a speech recognition system. Where some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation machines). The different classifications of the types of vocabularies in Automatic Speech Recognition systems (ASR) are as follows.

- Small vocabulary - tens of words
- Medium vocabulary - hundreds of words
- Large vocabulary - thousands of words
- Very-large vocabulary - tens of thousands of words
- Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word
The environment variability, channel variability, speaking style, sex, age, speed of speech also makes the Automatic Speech Recognition systems (ASR) more complex apart from the above characteristics. But the efficient systems must cope with the variability in the signal.

1.6 Applications of speech recognition

Today, speech recognition systems find widespread application in tasks that requires human machine interface. Speech recognition can be applied in applications such as railway reservations, query based information systems that provide updated travel information, weather reports, access to information: travel, banking, Commands, Avionics, Automobile portal, speech transcription, voice dictation, handicapped people (blind people) supermarket, stock price quotations, automatic call processing in telephone networks, data entry etc. Within telephone networks, speech recognition technology was increasingly used in automation as well as to enhancing the operator services.

<table>
<thead>
<tr>
<th>PROBLEM DOMAIN</th>
<th>APPLICATION</th>
<th>INPUT PATTERN</th>
<th>PATTERN CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Communication Sector</td>
<td>Telephone directory enquiry</td>
<td>Speech waveform</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Education Sector</td>
<td>Teaching students of foreign languages to pronounce vocabulary correctly.</td>
<td>Speech waveform</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Dictation</td>
<td>Dictation systems on the market accepts continuous speech input which replaces menu system.</td>
<td>Speech waveform</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Sector</td>
<td>Description</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Translation</td>
<td>Application which translates from one language to another</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Medical sector</td>
<td>Health care, Medical Transcriptions (digital speech to text)</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>General</td>
<td>Automated transcription, Telematics, Air traffic control, Multimodal interacting, court reporting, Grocery shops</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Differently abled</td>
<td>Useful to the people with limited mobility in their arms and hands or for those with sight</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Artificial Intelligence sector</td>
<td>Robotics</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Multimedia sector</td>
<td>Interactive video games and Gambling</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Domestic sector</td>
<td>Control of appliances such as Fridge, Owen etc.</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
<tr>
<td>Military field</td>
<td>Ware-fare management, air traffic controller training</td>
<td>Speech wave form</td>
<td>Spoken words</td>
</tr>
</tbody>
</table>
1.7 Proposed system in the Research Work

An intelligent system for Tamil speech recognition is proposed in this Research work. Recognition of Tamil speech and Tamil numerals is carried out in this proposed system. Thus various steps involved in speech recognition can be summarized as follows

- **Speech Signal acquisition:** From a digitized microphone or telephone, a raw Tamil speech signal is received with high frequency which is about 8 Hz over a telephone or 16Hz over a microphone. In-order to perform further processing by computer, the speech signal obtained from microphone which is initially in the analog form needs to be converted to digital signal. This process of conversion of signal is carried out by Analog to Digital (A/D) converter.

- **Feature Extraction:** In the proposed work, MFFC is used to carry out the Feature extraction process, which is one of the widely used techniques for the purpose. In-order to provide a compact representation of the given input signal, a sequence of feature vectors is computed during Feature extraction, which is a very important phase of intelligent Speech recognition development. Speech signal undergoes speech analysis which acts as first stage of Feature extraction process where the envelope of power spectrum described by raw features, are generated. In the second stage an extended feature vector composed of static and dynamic features is compiled. Finally transformation of this feature vector into more compact and robust vector is done.
• **Feature matching**: The process of feature matching is carried out using that of Dynamic Time Warping (DTW). Feature matching is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. The feature matching process starts with the test data being converted to that of templates. The feature matching process then consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon dynamic programming.
1.8 Thesis Outline

The gist of each chapter in the thesis is provided here under.
Chapter 2 describes Literature Survey on the field of speech recognition. Various speech recognition tasks are described: speaker-dependent vs speaker-independent, isolated words vs continuous speech, small vocabulary vs large vocabulary. The constraint and difficulties of each task is discussed. The different approaches applied to speech recognition are described. Next described are the various classification and modeling techniques in speech recognition. Their relative strengths and weaknesses are identified. From the review of different approach and techniques, the most suitable one are adapted, thus the scope of developments are identified.

Then, the current profound speech recognizers and their performance are presented. The current Tamil speech recognizers are reviewed. From the review, their limitation and constraints are identified: speaker dependent, isolated words, small vocabulary. Based on the review, derived the objective of the thesis to solve these constraints and provide a basis in developing speaker independent Tamil large vocabulary continuous speech recognition system.

Chapter 3 describes about the Feature Extraction concepts using Mel Frequency Cepstral Coefficient (MFCC). Initially in the chapter purpose of feature extraction is discussed. Later the various steps involved in MFCC is described.

Chapter 4 describes the Feature matching concept using DTW. In this chapter initially the basic concept of feature matching is described. Further it discusses about the concepts in Dynamic Time Warping (DTW), the DTW algorithm. It also discusses about the recognition of isolated words and Continuous words.

Chapter 5 gives the Performance analysis of speaker dependent, speaker independent, isolated Tamil word, Tamil numeral, Tamil connected word and Tamil continuous speech
recognition system in noiseless and noisy environment. The recognition accuracy of the individual Tamil words and Tamil numerals are also analyzed. The noise performance of the system is analyzed by considering the different Peak Signal to Noise Ratio (PSNR) values and the recognition accuracy is summarized.

Chapter 6 concludes the thesis with the key findings, where we summarize our contributions and provide possible future research directions.