CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

The production of artificial speech has been a cherished dream of mankind for quite a long time. Emotional Speech Recognition and synthesis has emerged as an important research area in the recent past. This is because natural artificial speech communication is a long cherished dream and will be highly useful for human beings [47]. There has been a great achievement in the development of artificial speech synthesis system that has high accuracy and intelligibility. But artificially produced speech lacks the naturalness of human speech. It is very important that such artificial systems should have the ability to process non-linguistic information such as emotion so as to convey the proper message. When the component of emotion processing is embedded into test-to-speech systems, it becomes more effective and natural. Hence artificially produced emotional speech synthesis provides a lot of naturalness to synthetic speech. The recognition of emotions is the key factor for emotional computing. The knowledge about the feature vectors that contain the emotions is important for interpreting the semantic meaning of the signal.

2.2 Text-to-Speech System

A basic **Text to Speech** (TTS) Synthesis system for emotional speech consists of two phases. They are as follows:

(a) **Text analysis**: The first phase is referred as the text analysis phase. It is the process by which the text input is transcribed into its linguistic form of representation. It is also known as high level synthesis.

(b) **Generation of speech waveforms**: This is the second phase in which the acoustic output is generated from all the prosodic and phonetic information generated in the text analysis phase. This phase is also known as **low-level synthesis**.

2.3 Architecture – Text-To Speech (TTS) System

A TTS system is a complex system since it is the culmination of digital signal processing, language processing as well as computer science [68]. The main approach that is taken in building a TTS system is to extract the acoustic and phonetic characteristics that are incorporated in the speech signal and to make use of it along with the linguistic information to generate the correct speech. The quality of a TTS system is evaluated in terms of features stated below [67]:

(i)The level of accuracy for input text analysis.

(ii) The level of intelligence of the resultant speech signal and

(iii) The level of naturalness of the resulting speech.

The basic TTS system mainly consists of two main modules:-

(i) **Natural Language Processing (NLP) module**: The high-level processing of text is done in this phase. It consists of linguistic analysis, phonetic analysis, text normalization, etc.

(ii) **Digital Signal Processing (DSP) module**: It performs the low-level process of the speech waveform generation [45].

The functional diagram of a TTS Synthesizer is shown below:

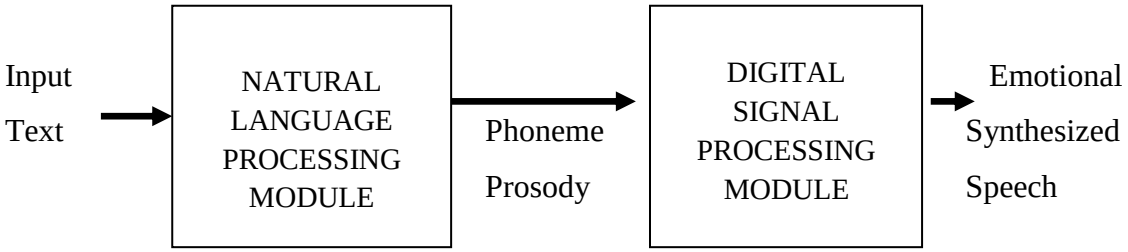

Fig 2.1: Diagram showing a general TTS System

2.3.1 TTS- Natural Language Processing (NLP)

In this stage, the input text is syntactically processed to generate a phonetic transcription. Here, the sentences present in the text are broken up into words by using regular grammars. Also the numbers, acronyms and abbreviations which are there in the text are also expanded. The output of the Natural Language processing module (NLP) is a phonetic transcription corresponding to the input text. This phonetic transcription also contains the desired intonation and prosody [69]. The architecture of a NLP module with its different components is shown in Fig 2.2.

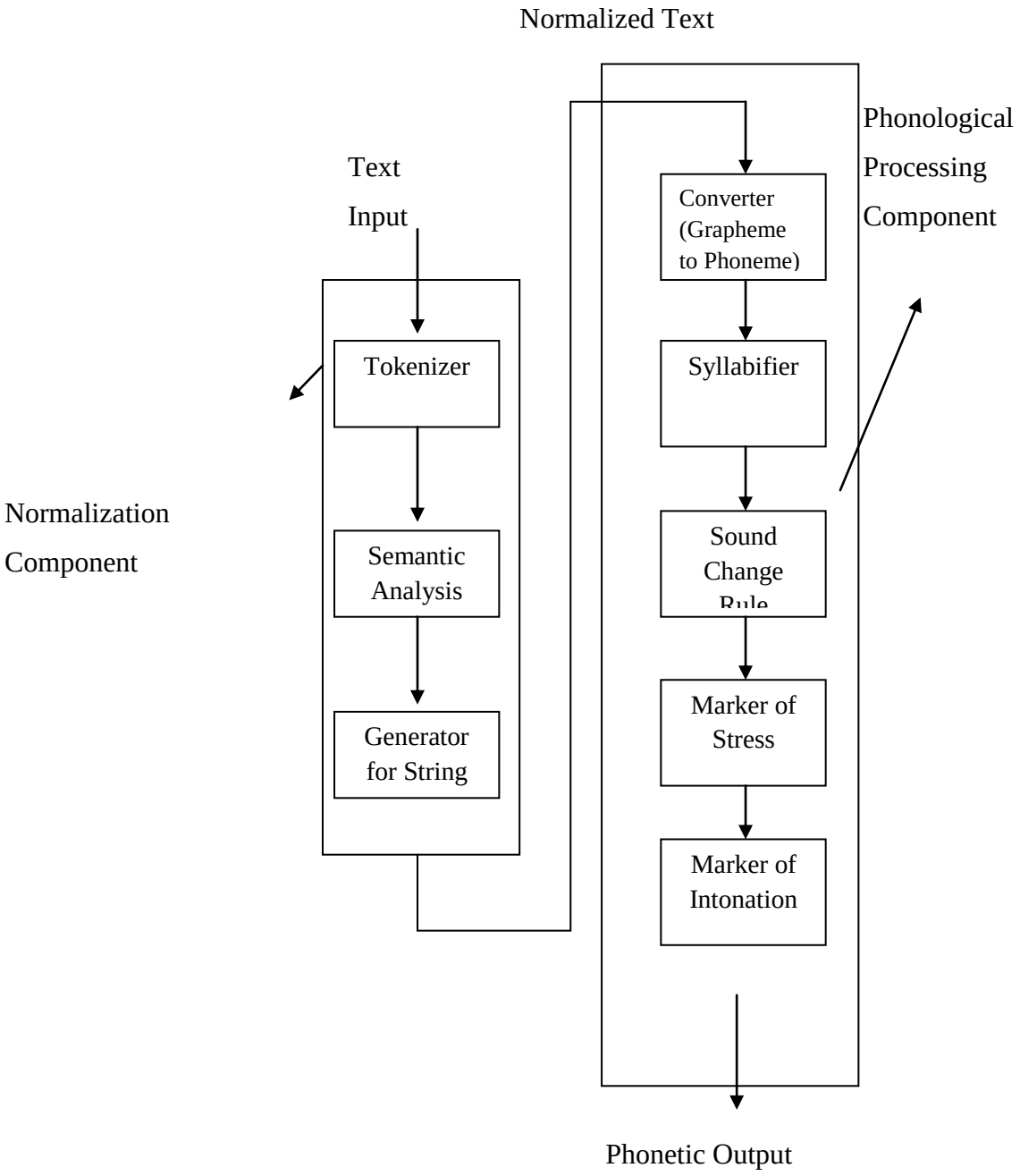

Fig 2.2: Architecture of a NLP module with its different components

The Natural Language Processing (NLP) module mainly consists of two components:-

(I)**Text normalization component**: Syntactic processing has to be carried out for the text input to generate the phonetic transcription. This component accepts a

character string as the input processes it and finally transforms it into a string of letters. This module consists of a Tokenizer which is used to mark the punctuation and space between words so that the token boundaries are marked. The output of the Tokenizer may be the words, date, time punctuations etc.

It also consists of a Semantic Tagger where the sentences of the input text are divided into fragments like words, numbers, etc. Also, the abbreviations and acronyms present in the text are expanded. This is accomplished by using regular grammars. Another component i.e the String generator accepts any non-letter based input, like a date containing digits and then converts the same into a letter string.

(II) **Phonological processing component**:

The Phonological Processing Component takes as input a string of letters and converts it into its corresponding phonetic transcription. This process is carried out with the help of various sub-modules. Here, the first step is the grapheme-to-phoneme conversion which takes the normalized text as input and converts it to a phonemic string. This process is also known as the **letter-to-sound** conversion. The sound generated then passes through a Syllabifier, which helps to marks syllable boundaries. This output is then passed through a processor which applies some sound change rules to produce the corresponding phonetic string. The Prediction of the correct prosodic characteristics from the plain text is a difficult task. The prosodic features also depend on many different aspects like the speaker's gender and age, emotions contained as well as the type of the sentence [70]. This task is carried out by generating the syntactic as well as the prosodic structure of the sentences analogous to the clauses and phrases. The durations of these speech units as well as the intonation to be used is applied on these

units. This step utilizes phonetic and phonological knowledge acquired from experts or through statistical methods like CART [71]. The characteristic like stress and intonation are added to the string by the modules Stress Marker and Intonation marker. Lastly, the annotated phonetic output generated from Text Normalization component is given as input to the Digital Signal Processing (DSP) module which produces the synthetic speech [72].

2.3.2. Digital Signal Processing (DSP) Module

The DSP module acts like the human speech production system to generate the synthesized speech [101]. The input to the DSP module is the phonetic transcription containing all the prosodic information. The technique chosen by DSP to produce the synthesized speech depends on many factors like the language of the input text, platform used and the motive of building the system. The quality of the synthesized speech should be good enough for understanding the context in which it has been spoken even though it is very difficult to produce natural speech like humans. The common techniques for speech synthesis used by the Digital Signal Processing (DSP) module have already been discussed in chapter-1.

2.4 Evaluating Text-To-Speech Systems Performance

The performance of a TTS System needs to be evaluated to achieve the purpose for which it has been built. As mentioned by Klatt [74], there are some criteria for the evaluation of a TTS System like intelligibility of individual **phonemes**, **words** / **words in naturalness**, **context of sentence**, and **suitability** of the system for a particular application [73]. Some of the criteria for evaluation are discussed below:

(i) **Intelligence based on isolated words:** There are various ways to evaluate the intelligence based on isolated words of synthesized speech. A method called the modified rhyme test is often used for consonants which are found to be more difficult to synthesize as compared to vowels. In the test, the listener selects among six familiar words which is different only by one consonant, i.e an initial consonant or the final consonant. This method to evaluate the performance of the system is not very severe. This is because the alternative response may leave out a confusion that would be created if a blank answer sheet is being used. But the test does make it possible the rapid presentation of naive subjects like the automatic scoring of answer sheets.

(ii) **Intelligence for words in sentences:** The words spoken in a language have significant co-articulation in terms of word boundaries, phonetic simplifications, prosodic modifications and reduction of unstressed syllables, shorten non final syllables and also modify the fundamental frequency contour. The tests of word intelligence in sentence frames have been devised to evaluate the performance of the TTS system based on these transformations.

(iii) **Naturalness of synthesized speech:** Naturalness of the synthesized speech is an attribute which is not easy to quantify because it is highly subjective. There are a large number of deficiencies that can occur and make synthesized speech sound unnatural to a large extent. However, TTS systems can be compared and evaluated for subjective naturalness. This can be done with a high degree of subject-wise and test-retest agreement.

A standard practice has been to play pairs of test sentences which are synthesized by each of the systems to be compared, and acquire the judgments. For

doing this, the sentences to be compared have to be the same, and they have to be played without long pauses in between, and subsequently valid data can be obtained for the tests.

(iv) **Suitability:** The sustainability of the TTS system for a particular application has to be evaluated for its optimum performance. In recent times, Text-to-Speech systems are being used in devices for a large number of applications. These devices will be accepted the general public only if it is useful. For example, if the TTS system is used as part of some application that offers a new and innovative services like providing direct access to stored information on a computer, or allows cheap and easy access to a presently existing service etc.

2.5 Speech Synthesis- Historical Developments

Mankind has had a long cherished dream of natural artificial speech having emotions. This has led to tremendous developments in the field emotional speech synthesis.

2.5.1 Speech Synthesis - Early Attempts

The first device to be used as a speech synthesizer was introduced in 1939 by Homer Dudley and was known as Voice Operating Demonstrator (VODER)[55]. The quality and naturalness of speech produced by this synthesizer was not very good. George Rosen in 1958 introduced the first articulatory speech synthesizer at the Massachusetts Institute of Technology (M.I.T) [48]. In 1952 a **Pattern Playback synthesizer** was developed by Franklin Cooper and his associates at the Haskins Laboratories [49]. In 1953, Parametric Artificial Talker (PAT) was introduced by

Walter Lawrence [50]. It was the first formant synthesizer for synthesizing artificial speech. In 1773, Christian Kratzenstein, a Russian Professor developed an apparatus to generate the five long vowels (/a/, /e/, /i/, /o/, and /u/) artificially. In 1791, Wolfgang von Kempelen introduced an "**Acoustic-Mechanical Speech Machine**", which was used to generate single sounds and some sound combinations [51]. In the 1800 century, Charles Wheatstone constructed a speaking machine which was a variation of von Kempelen's acoustic machine. In late 1800 Alexander Graham Bell [52] along with his father introduced a similar kind of speaking machine as Wheatstone.

2.5.2 Achievements in Speech Technology during 1960's

In 1960's new hardware techniques were developed with specialized purposes which could recognize speech [53]. In 1963, a vowel recognizer was developed at the Radio Research Lab in Tokyo by J.Suzuki and K. Nagata [54]. A recognizer for phonemes was built by J.Sakai and S.Doshita at the Kyoto University in 1963 [56]. This can be considered as the starting point of recognition of continuous speech. The most prominent development of that time was the digit recognizer built at NEC Laboratories by K.Nagata , S.Chiba and Y.Kato in 1963 [54]. The analysis on speech segments from different parts of the speech signal for recognition purposes was first shown by J.Sakai and S.Doshita. A phoneme recognizer which was capable of recognizing four vowels and some consonants was developed at the University College of England by D.B Fry and P Denes in 1959 [57]. This was further developed to add statistical information to recognize phoneme sequence for the English language. The concept of utilizing a speech segmenter with the help of a non-uniform time scale for the alignment of speech patterns was used for development at RCA Laboratories in 1960 by Tom Martin and T.K Vintsyuk [100]. A method for detecting the endpoint of the utterance was proposed

by Martin. This led to the enhancement of performance of Speech recognizers [57]. Also, a technique for meaningful assessment for time alignment between two utterances by using dynamic programming was initiated by T.K Vintsyuk [100]. In the middle of 1960's **Linear Predictive Coding** (LPC) came into limelight. This made it easier to estimate the vocal tract response from the speech signal. The first fully developed text-to-artificial speech system for the English language was developed in the Electrotehnical Laboratory, Japan 1968 by Noriko Umeda and his companions [74].

2.5.3 Achievements in Speech Technology during 1970's and 1980's

During 1970's, many commercial versions of speech synthesizer were introduced in the market [74]. The first integrated circuit used for speech synthesis was the **Vortrax chip**. In 1978, a Vortrax-based Type-n-Talk system was introduced by Richard Gagnon was introduced which was quite inexpensive. Sometimes later, inexpensive linear prediction synthesis chip (**TMS-5100**) was used to develop a **Speak-n-Spell synthesizer,** based on LPC. It received great attention as it was used as an electronic aid for reading. Moreover, new versions of the synthesis chip **TMS-5100**, i.e **TMS-5220** was developed in 1982 by Street Electronics to make low cost diphone synthesizers like "Echo".

In the middle of 1970's, the concept of using fundamental pattern recognition techniques to speech recognition area, based on LPC was proposed by L.R.Rabiner, S.E Levinson [103] and F.Itakuru [104]. It was during this time that Tom Martin launched his commercial company "Treshold Technology Inc" for speech recognition [75].The first automatic speech recognizer by the name "**VIP-100 system**" was launched by this company. This was used in very few applications like quality control, package sorting

etc. However, this work inspired the Advanced Research Project Agency from U.S Department of Defense which resulted in the establishment of "Speech Understanding Research" Program. At the same time, **IBM** and **Bell Laboratories** initiated the research on Speech Recognition. A **Voice Activated Typewriter (VAT)** which had the ability to convert uttered sentences into a sequence of letters and words in different formats was developed at the Bell Laboratories [102].The main drawback of this system was that it was highly speaker dependent and had to be trained for every new speaker. It was also dependent on the size of the vocabulary and also the structure of the language.

The field of Telecommunication saw a landmark initiative in the form of an **Interactive Voice Response (IVR)** system which was initiated at the AT&T Bell laboratories. The main motive of developing this system was to build a speaker independent system which could work with variable speech signals from different speakers with different accent. H. Dudley, S.A Watkins and R.R Riesz developed many Speech clustering algorithms for word creation and referencing sound patterns to be used for many speakers with different accent. Out of these the **Itakaru distance and statistical modeling technique** generated high quality representations of the speech signals for many speakers [76-77]. The AT&T Bell laboratories also laid importance on the spectral representation of speech signals. Another significant step of the Bell laboratories was the introduction of keyword spotting which was used as a basic form of speech understanding [105]. In 1979, a text to speech system was developed at M.I.T laboratory, by the name MITalk by Klatt and Hunnicutt. After some time with few changes, the system was launched commercially by TSI (Telesensory Systems Inc.) [29]. Later, Dennis Klatt invented the Klattalk system, which utilized a sophisticated voicing source and became very famous, [37]. The technology used behind MITalk and

Klattalk systems form the ground for recent developments in many speech synthesis systems today. The development of the statistical methods especially the HMM framework in 1980's has become the basis of most modern speech recognition systems

2.5.4 Recent Developments

In recent times, technologies used for Speech synthesis is gradually becoming highly complicated and more sophisticated. The Hidden Markov model (HMM) is more in use in recent times and is becoming quite popular [80]. The HMM technique has been applied to Speech synthesis / recognition since late 1970's. Hidden Markov model (HMM) based speech synthesizer/ recognizer are most commonly used methods in recent times. The Hidden Markov model Toolkit (HTK) was developed by Steve Young at the speech vision and Robotics group of the Cambridge University Engineering Department in 1989.

The Hidden Markov model is a compilation of states connected by transitions which has two sets of probabilities in each: (i) **transition probability –** It is the probability of whether the transition will be taken or not and (ii) **output probability density** function (pdf) – It is the conditional probability of a finite alphabet generating an output symbol. The field of speech synthesis/recognition has seen the application of Artificial Neural Networks (ANN) in the recent years. ANN has also shown positive results in speech synthesis in recent times.

In 1977, the **SYNTE2** was introduced, which can be considered the first properly developed speech synthesizer [44]. About a few years, an improved version, **SYNTE3** synthesizer was launched in Finland and it was a market leader for quite a long time.

In 1980's, several other commercial systems were introduced which used a chip called the **Votrax speech synthesis** chip. Examples of such systems were the Caiku, Eke, Seppo, Amertronics, etc. Two synthesizers, which were based on concatenation - **Mikropuhe** and **Sanosse**, are probably some of the best known products in this field. Sanosse synthesizer was developed for educational use for the University of Turku. However, the best known synthesizer is perhaps the **festival synthesizer** which was developed in Sweden.

2.6 An Overview of the Proposed Approach

In the present work, the main methodology that has been followed is the analysis of the acoustic and phonetic properties of emotional Bodo speech signal and its use for the development of Bodo emotional speech synthesis and recognition system. Here, the Fast Fourier Transform (FFT) is mainly used to obtain the frequencies from input speech signal. FFT transforms the continuous speech time domain signal to its corresponding frequency domain signal. The noise component of the signal is discarded with the help of a filter called the Finite Impulse Response (FIR). Linear Predictive Coding (LPC) is made use of to detect the locations of the formats. This is because the LPC coefficients give stress on the location of the formants in the frequency spectrum. A vector of linear predictive coefficients is generated which is shown by a smooth spectral envelope representing the FFT magnitudes. The location of the formants, which is very important for analysis of the speech signal is thus shown spectrally. The LPC coefficients help to determine and generalize the unique characteristics of the speech signal as well as the speaker. The Mel frequency Cesptral Co-efficient(MFCC), Short Term Energy(STE), Zero crossing Rate(ZCR), Hidden Markav Model(HMM) are all utilized to analyze the features of Bodo phonemes, words etc. The three formant

frequencies i.e., F1, F2 and F3 for each Bodo Vowel and some Bodo words of CV(consonant-vowel), VC(vowel-consonant) and CVC(consonant-vowel-consonant) type word to determined to find the distinctive differences between the three emotional states under consideration i.e normal, angry and surprise moods. A Study to investigate the cepstral measures to differentiate between the two different genders of speakers, i.e male and female, for the Bodo Language is carried out. The MFCC and HMM methods are used to analyze the feature of Bodo phonemes.