CHAPTER 2

Literature Review

2.1 Introduction

Synthesis of speech and automatic generation of speech wave form is under study for the last few decades [14, 15]. Although there have been tremendous development had happen and many has produced synthesizers with a high degree of accuracy and intelligibility, but still it lacks to produce better quality sound and naturalness. As far as TTS synthesis process in concerned, it contains two main phases. The first phase, also known as **High Level Synthesis** associated with the linguistic representation of input text, which is known as **Phonetic Representation**. The second phase also known as **Low Level Synthesis** is associated with the production of acoustic output based on the phonetic and prosodic information collected in the first phase. The entire setup is shown in the figure.2.1.

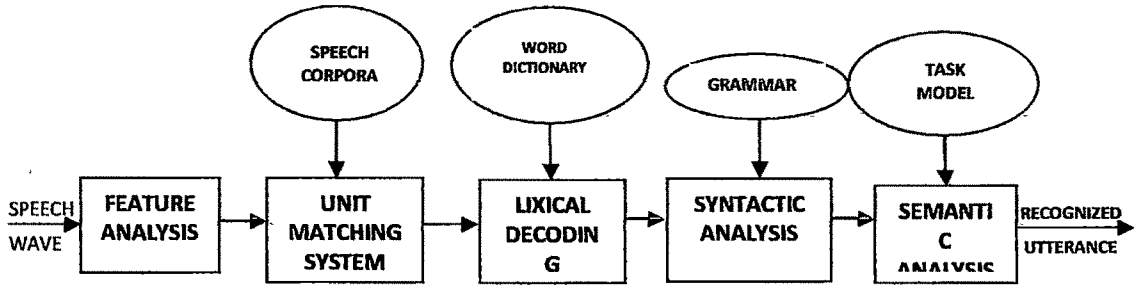

Figure 2.1 : Speech Synthesis System

One of the simplest ways to produce synthetic speech is to play long pre-recorded samples of natural speech such as single words or sentences. This method produces high quality and naturalness. However, it has got the limitation of limited vocabulary and limited number of informants. The method seems to be suitable for some announcing and information system. On the other hand it is very difficult to create a database of all words of a particular language in a recorded form. Even if it is possible, it will not be justified to call it as Speech Synthesis system. Thus to design a system with greater flexibility, we need to consider shorter pieces of speech signals, such as syllables, phonemes, diphones, or even shorter segments.

Another widely used method to produce synthetic speech is formant synthesis, which is based on the **source-filter-model** of speech production described in Figure 2.2.

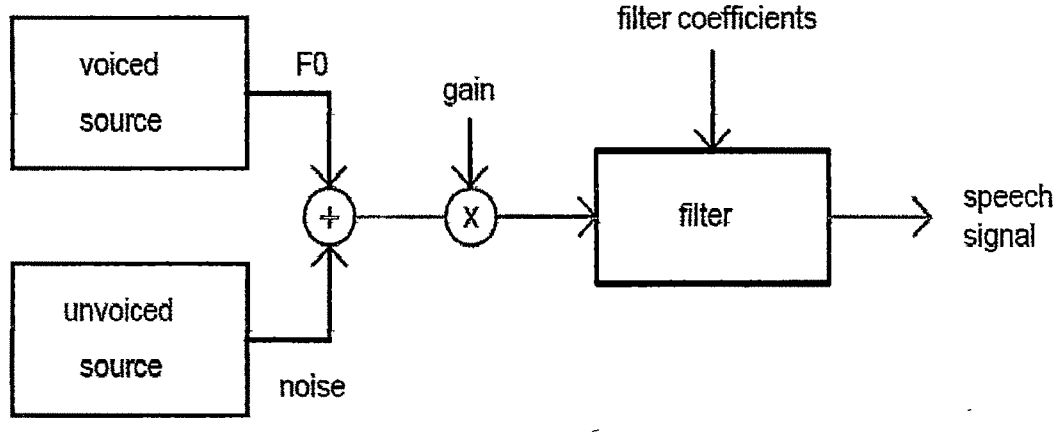

Figure. 2.2 : source-filter-model of speech production

This method sometimes called as **Terminal Analogy** because it models only the sound source and the formant frequencies, not any physical characteristics of the vocal tract [16]. The excitation signal could be either voiced with

fundamental frequency (F0) or it could be unvoiced noise. A mixed excitation of this two may also be used for voiced consonants and some aspiration sound. The excitation is then amplified and filtered with a vocal tract filter which is consisted of resonators similar to the formants of natural speech.

Theoretically, the most accurate method to generate artificial speech is to model the human speech production system [17, 18, 19]. This method known as **Articulatory Synthesis.** It involves models of the human articulators and vocal cords. Although method produces high quality Synthetic speech, but due to its complexity, its potentiality is not yet been realized.

All synthesis methods have some merits and de-merits of their own and it is quite difficult to say, which method is most appropriate one. With concatenative [19] and formant synthesis [17], very promising results have been achieved [17]. But in the time to come, the articulatory synthesis may arise as a potential method in the future [17].

2.2 Different Approaches to Speech Recognition

In this section, I tried to discuss the different approaches to speech recognition, which are so-far used by the different group of researchers worldwide. I also tried to give a comparative study of the different techniques. Speech recognition deals with the recognition of the specific individual speech sounds [83]. Some of the commonly used techniques are – **Dynamic Time Wrapping (DTW)**, **Hidden-Markov Model (HMM)**, and **Artificial Neural Network (ANM)** etc.

2.3 History of Speech Recognition

Typewriters and computers in the name of DTP are being used for quite a long time to create a document. Over the last few decades, scientists and speech researchers have been working to make it a reality, that a computer could recognize the speech.

The attempt for developing a machine to mimic human's speech communication capability started later 18th century. The attempt was not emphasized on recognizing and understanding speech, but the interest was to develop a speaking machine .

2.3.1 Early Speech Production and Representation Technologies

In 1773, Christian Kratzenstein, a Russian scientist and professor of Physiology in Copenhagen, succeeded to produce vowel sounds using resonance tubes connected to organ pipes [20]. Later in 1791, Wolfgang von Kempelen in Vienna constructed an "Acoustic–Mechanical Speech machine" [21]. In the mid 1800's, a version of Von Kempelen's speaking machine using resonator made of leather was built by Charles WheatStone [22].

During the first half of the 20th century, Harvery Fletcher [23] at Bell Laboratories documented the relationship between a given speech spectrum. In the 1930's, Homer Dudley developed a speech synthesizer called the VODER [21].

In the early attempts of the designing of system for automatic speech recognition, the theory of acoustic phonetic played the vital role which describes

the phonetic elements of speech and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic context. For example, to produce a steady vowel sound, the vocal cord needs to vibrate and the air that propagates through the vocal tract results in sound with natural modes of resonance, called the **Formants** or **Formant Frequency**. They considered as the major region of energy concentration in the speech power spectrum.

Later in 1952, a system for isolated digit recognition for a single speaker was developed by Davis, Biddulph and Balashek of Bell Laboratories [24]. They used the Format Frequencies measured during vowel regions, of each digit. The plots of the formant frequencies along the direction of the first and the second formant frequencies served as the "reference pattern" for determining the identity of an unknown digit utterance.

Earlier, during 1950's Olson and Belar of RCA Laboratories built a system to recognize 10 syllables of a single speaker [25] and at MIT Lincbln Lab, Forgie and Forgie, built a speaker independent 10 vowel recognizer [26].

2.3.2 Development took place during 1960's

During 1960's, some special purpose hardware techniques were developed in Japan to perform the tasks of speech recognition [29]. In this connection, we can name J. Suzuki and K. Nagata (1963) for the development of the vowel recognizer t the Radio Research Lab in Tokyo [27]. Another Phoneme recognizer was also

developed at Kyoto University by J. Sakai and S. Doshita (1963) [28]. But out of all, the digit recognizer of NEC Laboratories designed by K. Nagata, Y. Kato and S. Chiba (1963) [29] can be considered as the most prominent one. Later Sakai and Doshita introduced the use of a speech segmenter for analysis and recognition of speech in different portions of the input utterance for the first time. In contrast, the isolated digit recognizer does not need a segmenter where it was assumed that the unknown utterance contained a complete digit. The work of Kyoto University could be considered as the point of inception as far as the continuous speech recognition is considered.

D.B. Fry and P Denes (1959) developed a phoneme recognizer to recognize four vowels and some consonants at University College in England [30]. Later on, they incorporated statistical information to find out allowable phoneme sequences in English. The use of statistical syntax at phoneme level for the development of automatic speech recognition was also introduced in this work by them.

Tom Martin (1960) at RCA Laboratories [31] and T.K. Vintsyuk (1960) in the Soviet Union, introduced the concept of using a speech segmenter by adapting a non-uniform time-scale for aligning speech patterns [32]. Martin [31] also proposed a method for determining the utterance endpoint, which greatly enhanced the reliability of the recognizer performance [31]. T.K. Vintsyku (1968), proposed a technique using dynamic programming to derive a meaningful assessment for time alignment between two utterances [32].

2.3.3 Development of speech Recognition systems in 1970's

Followed by the work of Vinksyuk (1960), another more formal method, generally known as **Dynamic Time Wrapping (DTW),** was proposed by H. Sakoe and S. Chiba [33] to recognize the speech pattern. Consequently, some improvement took place, where **Viterbi Algorithm** was included to the Dynamic Programming producing variant forms and finally becomes a vital method in automatic speech recognition, which was supported by the Publication of Sakoe and Chiba [34].

In the late 1960's, the fundamental concept of **Linear Predictive Coding (LPC)** was formulated independently by B. S. Atal and F. Itakura [35, 36]. This concept significantly simplified the estimation of the vocal tract response from speech waveforms.

During mid 1970's, F. Itakura[37]; L.R. Rabiner and S.E. Levinson [38] proposed the concept of applying fundamental pattern recognition technology to speech recognition based on LPC methods. During the same time Tom Martin(1960) [31] established the first speech recognition commercial company Threshold Technology Inc. and developed the first real ASR product called the **VIP-100 system.**

The VIP – 100 System was used in a very few simple applications such as by television faceplate manufacturing firms (for quality control) and by **FedEx** (for package sorting). The work influenced the **Advanced Research Project Agency (ARPA)** of the U.S. Department of Defense, resulting in establishment of **Speech**

Understanding Research (SUR) program. During the same time Speech Recognition Research was also initiated at **IBM** and **Bell Laboratories**.

At IBM laboratory, a **"Voice Activated Typewriter" (VAT)** was planned to be developed, which could convert the uttered sentences into sequence of letters and words, which could be displayed in different formats [41]. The basic drawback of the system was that it was totally speaker dependent, need to be trained for each new speaker. It was focused on the size of the recognition vocabulary and the structure of the language model i.e its grammar. It depicted the emergence of sequences of language symbols like phonemes or words in the speech signal in a probabilistic sense represented by statistical syntactical rules. This type of speech recognition tasks are commonly known as **transcription** and the set of statistical grammatical or syntactical rules are commonly known as **Language model**. Thus the **n-gram** model defined the probability of occurrence of an ordered sequence of **n-words**.

Another initiative taken at AT&T Bell Laboratories research program in the field of Telecommunication was- **Interactive Voice Response (IVR)**, **voice dialing** and **command and control routing** of phone calls. The main objective of this research was to design a speaker independent system, which could deal with the acoustic variability which is intrinsic in the speech signal of different speaker with different accent. A range of speech clustering algorithms for creating word and sound reference patterns to use across a wide range of speaker and accent were developed by H. Dudley, R.R. Riesz, and S.A. Watkins [21]. Among these the

Itakura distance [42] and **statistical modeling techniques** [43] can be mentioned for generating sufficiently rich representations of the utterances from a vast population. The work at Bell Laboratories also conferred stress on the spectral representation on sounds or words generally called the Acoustic Model over the representation of the grammar or syntax of the task called the **Language Model**. The concept of Keyword spotting as a primitive form of speech understanding was another importance of the approach of Bell Laboratories [43]. Though different realizations of the technology raised due to the different in goals in various applications, a certain degree of convergence in the system design acquired with the rapid development of statistical methods in the 1980's, among which the most importantly the **Hidden Markov Model (HMM)** framework [44, 45]. The statistical framework has become the basis of the most practical speech recognition systems in the recent time with some significant development and modification to the system.

2.3.4 Review of the work in the remaining years

During 1980's, the speech recognition technology shifted towards the pattern recognition approach supported by the statistical methods. In a statistical approach, inventories of elementary probabilistic models of basic linguistic units like phonemes are used to build word representations. A set of acoustic parameters, extracted from a spoken utterance, is seen as a realization of a concatenation of elementary process described by **Hidden Markov Model (HMM)**. The HMM for speech recognition got the popularity after the publication by S.E. Levinson, L.R.

Rabiner, M.M. Sondhi and J.D. [45, 46]. The famous HMM model is based on the years old mathematical model known as **Markov Chain**.

The model was first introduced by **Institute for Defense Analysis (IDA)** in Princeton. Later, with the increase in the use of the HMM technique by a large number of people, the constraint on the form of the density functions entailed a limitation on the performance of the system, particularly for speaker independent process. This is due to the fact that the speech parameter distribution was not sufficiently well modeled by a simple log-concave or an elliptically symmetric density function. In the early 1980's, at Bell Laboratories, it was established significantly to ensure the accuracy of recognition, particularly for speaker independent, large vocabulary speech recognition process by extending the theory of HMM to **mixture densities** [42, 47].

The combination of the **HMM** theory and the **Finite State Network** was an important technological development during the mid 1980's. The finite state graph is realized as a Markov chain for calculation of the likelihood, based on the observation sequence of an unknown utterance.

Artificial Neural Network (ANN) was another technology introduced in late 1980's. Although Neural Network was introduced during 1950's, initially it could not produce notable result [48]. The significance around the old initiative of mimicking the human neural processing mechanism was revived by the advent of a **Parallel Distribution Processing (PDP)** model. The PDP model was the association of simple computational elements, and a corresponding **"training"**

method, called **error back propagation**. Early attempts using Neural Network (NN) for speech recognition centered on simple tasks like recognizing a few phonemes or a few words (may be digit) with high success rate [44]. However, as the problem of speech recognition requires handling of temporal variation, neural networks in their original form have not proven to be extensible to these tasks.