CHAPTER-II

METHODS OF SPEECH SYNTHESIS AND ANALYSIS

The ultimate user-friendly machine, Robot or a personal computer, must have complete voice-communication abilities. Voice communication involves two separate but related technologies; speech synthesis and speech recognition. Speech synthesis is the more developed technology. The various methods of speech synthesis/analysis are as follows.

Electronic Speech Synthesis Technique:

Speech synthesis is the putting together, or creation of speech. Basically, two general techniques are used for electronic speech synthesis (ESS):

- Natural speech synthesis/analysis; and
- Artificial constructive/synthesis.

2.1 NATURAL SPEECH ANALYSIS/SYNTHESIS:

This techniques involves the recording and subsequent playback of human speech. The recording can be in analog form on magnetic tapes as well as in digital form by digitizing and storing in memory.

The analog record/playback method produces the most natural sounding speech, but it is not very practical for general speech production since the recorded massage must be accessed serially. In addition, it is almost impossible to record all the words and phrases. This is useful only for limited speech.

The digital record/playback method involves separate operations: analysis and synthesis. During the analysis phase, human speech is analyzed, coded into digital form and stored. Then during the synthesis operation, the digital speech is recalled from memory and converted to analog form to recreate the original speech waveform.

The digital analysis/synthesis method provides more flexibility than the analog method since the stored phrases can randomly accessed from computer memory. However, the vocabulary size is limited by the amount of memory available. For this reason, several different encoding techniques are used to analyze and compress speech waveform.

There are two types of digital analysis/synthesis method.

1. Time domain analysis/synthesis.
2. Frequency domain analysis/synthesis.

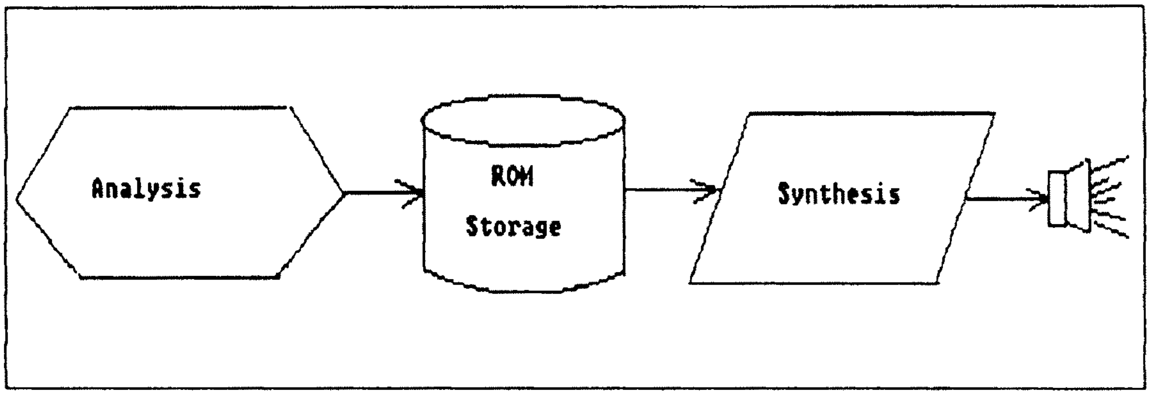


Fig. 2.1 Block diagram of Natural speech Analysis/Synthesis

2.2 TIME DOMAIN ANALYSIS/SYNTHESIS:

The speech waveform is digitized in the time domain, i.e. the analog waveform is periodically sampled and converted to a digital code using an A/D converter. The stored samples are then passed through a D/A converter to reproduce the speech during the synthesis operation.

2.3 FREQUENCY DOMAIN ANALYSIS/SYNTHESIS:

The frequency spectrum of the analog waveform is analyzed and coded. In addition, the synthesis operation attempts to emulate the human vocal tract electronically. Briefly, this is accomplished by using stored frequency parameters obtained during the analysis phase to control electronic frequency generators and filters that reproduce the voiced and unvoiced sounds of the human vocal tract.

In this technique a mathematical model of the frequency spectrum is stored and used to control an electric model of the human vocal tract.

During the analysis phase, the frequency characteristics of the human voice are analyzed to produce a series of mathematical parameters that are stored and subsequently recalled to control an electronic speech synthesizer. The electronic synthesizer emulates the human vocal tract using frequency generators and filters that are controlled by the stored speech parameters.

Two methods are employed for frequency domain analysis/synthesis.

1. Linear predictive coding.
2. Formant synthesis analysis.

2.3.1 Linear Predictive Coding:

It is commonly known as LPC. The first step in linear predictive coding is to digitize the speech waveform with an A/D converter using simple pulse code modulation. Once a digital form, the waveform is analyzed to extract the frequency intensity and other vocal tract type variables needed to mathematically reconstruct the waveform. The extracted speech data are then coded into a series of linear equation parameters called LPC codes, that models the frequency characteristics of the spoken waveform.

Once the speech waveform has been encoded into the LPC format, the stored speech parameters are used to control a synthesizer circuit.

The synthesizer circuit is designed as a model of the human vocal tract. It consists of three major sections, an excitation source, a multistage digital filter and D/A converter.

The excitation source includes a periodic pulse generator which emulates the action of your vocal cards by producing the periodic voiced sound frequency. The pitch of the sound is determined by rate at which vocal cards vibrate. Here, the frequency generated by periodic pulse generator determine the pitch of the synthesized sound. The excitation source includes a white noise generator, an electronic switch and an amplifier also. A white noise generator produces the unvoiced sounds. Actually, these sounds are not created by your vocal cards. It is produced due to air turbulence in the vocal cavity. Here, the white noise generator produces unvoiced sounds by generating a random frequency pattern. The combination of voiced and unvoiced sounds produces speech. This is the job of electronic switch. The voiced and unvoiced sounds are combined by electronically switching between the two sound generators. The selected sound is then amplified and passed through a multistage digital filter circuit.

The purpose of the digital filter is to shape or modulate. The excitation signal is modulated to produce the desired formant frequency spectrum by varying the characteristics of the filter, LPC data stored in ROM are used to control the filter characteristics. The digital filter output is then converted to an analog speech signal by D/A converter.

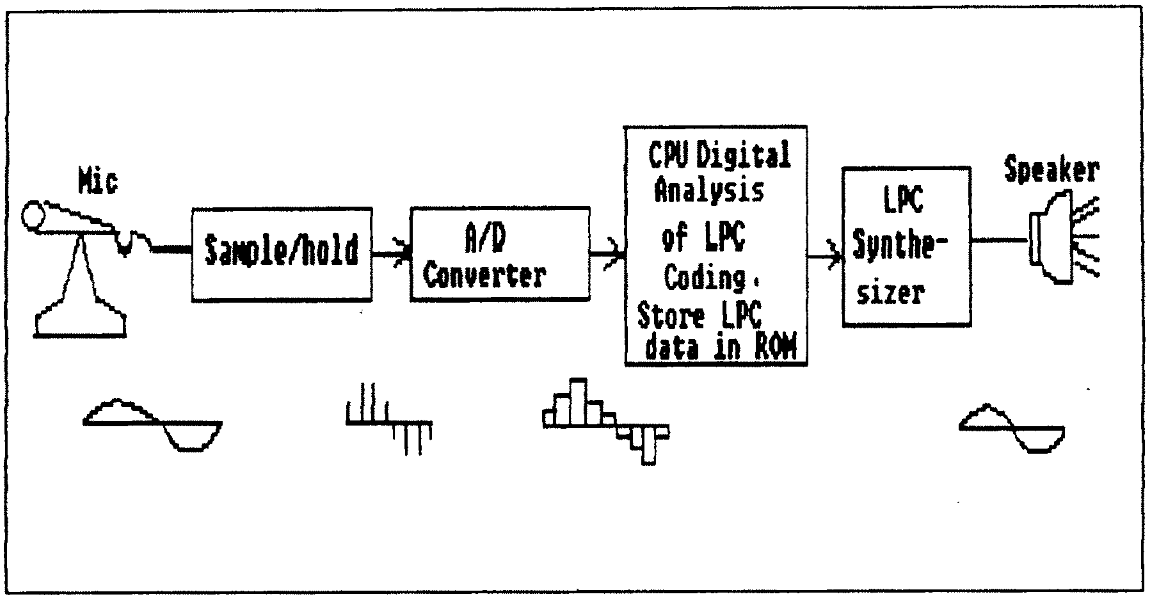
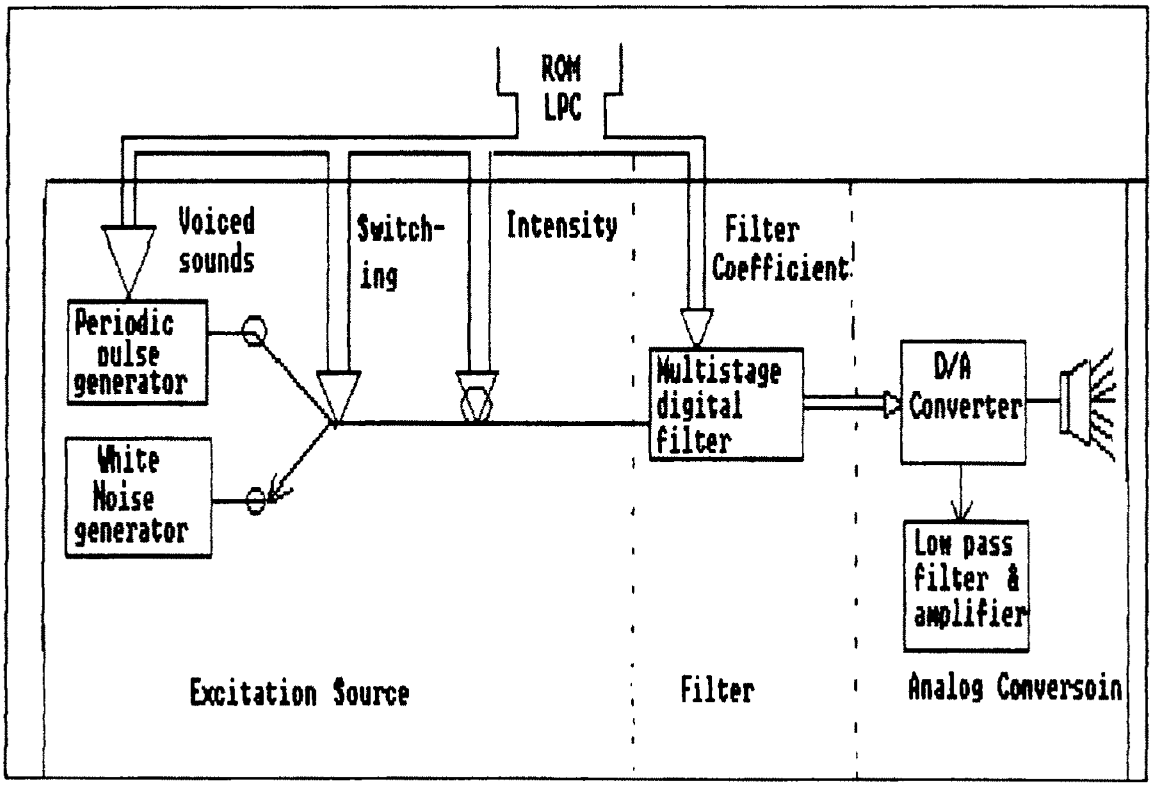


Fig. 2.2 Typical LPC system



(LPC Block Diagram)

Fig. 2.3 LPC speech synthesizer circuit models the human vocal tract

LPC data are stored in ROM. It controls the following functions:

1. Pitch of the voiced sounds;
2. Selection between the voiced or unvoiced sounds;
3. Amplitude of the excitation signal for sound intensity; and
4. Control of the digital filter by providing the filter coefficients required to modulate the excitation signal and produce the digital speech waveform.

LPC data rate required to reproduce speech is less than 2400 bps. The superior speech quality, coupled with the extremely efficient storage of speech has made LPC a very popular speech synthesis technique.

2.3.2 Formant Analysis/Synthesis:

It attempts to generate speech by reconstructing the formants that exists in the speech waveform.

The voiced sound consists of several resonant frequencies called formants. The formant frequencies are constantly, shifting to produce different sounds. The formant frequency characteristics of a spoken waveform can be digitally coded and used to control frequency generators and filters in an electronic synthesizer to reproduce the original speech waveform.

The digitized speech in formant analysis/synthesis can take on two different forms. The original speech formants can be coded and synthesized one word at a time. This is called the stored-word or dictionary. The disadvantage of this method is that the vocabulary is limited. Thus, to produce an allophone electronically, all you need to do is generate the unique formant frequency for that particular sound.

Phoneme speech synthesis is also a frequency technique that uses the principles common to both LPC and formant synthesis. Phoneme synthesis is a form of constructive analysis/synthesis.

2.4 ARTIFICIAL CONSTRUCTIVE SYNTHESIS:

Speech is created artificially by putting together or constructing the various sounds that are required to produce a given speech segment. The most popular method of doing this techniques is called phoneme speech synthesis. In this method phoneme and allophone sounds are coded and stored in memory. Software algorithms must then be written to connect the phonemes to produce a given word. Words are then strung together to produce phrases. In some cases the software algorithms must then be written to produce the rules that are used to translate written text into the appropriate allophone code. This is called as text-to-speech translation.

The artificial constructive synthesis is shown in Fig. 2.4.

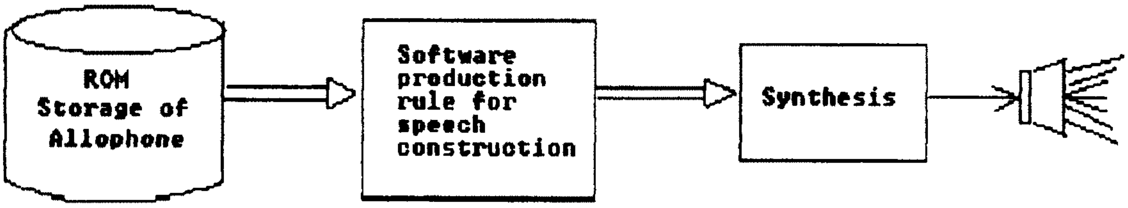


Fig. 2.4 Artificial constructive/synthesis (Block Diagram)

2.5 FREQUENCY ANALYSIS OF SIGNALS:

The signal representation basically involve the decomposition of the signals in terms of sinusoidal components or complex exponential. With such a decomposition, a signal is said to be represented in the frequency domain for the class of periodic signals, such a decomposition is called Fourier series. For the class of finite energy signals, the decomposition is called the Fourier transform.

2.5.1 Frequency Analysis for Continuous-Time Periodic Signals:

The basic mathematical representation of periodic signals is the Fourier series, which is linear, weighted sum of harmonically related sinusoids or complex exponential. A linear combination of harmonically related complex exponential of the form.

Synthesis Equation,

$$x\,(t) = \sum_{k=-\infty}^{\infty} C_k\, e^{j\,2\pi\,k\,f_o t} \qquad \text{...........}\ (2.1)$$

Where,

$x\,(t)$ = periodic signal

$Tp = \frac{1}{f_o}$ = Fundamental period

C_k = Fourier series coefficient

k = 0, 1, 2 ......

Analysis Equation,

$$C_k = \frac{1}{T_p} \int_{T_p} x(t)\; e^{-j\,2\pi\,k\,f_o t}\, dt \qquad \text{...........}(2.2)$$

In general Fourier coefficient are complex, so if signal is real,

$$C_k = |\,C_k\,|\, e^{j\theta k} \qquad \text{...........}(2.3)$$

and

$$C_{-k} = |\,C_k\,|\, e^{-j\theta k} \qquad \text{...........}(2.4)$$

∴ Fourier series may be represented in the form

$$x(t) = C_0 + 2 \sum_{k=1}^{\infty} | C_k | \text{Cos} (2\pi \, k \, f_o t + \theta k) \qquad .....(2.5)$$

Where C_0 is real valued and x (t) is real.

2.5.2 Frequency Analysis for Continuous time aperiodic signals:

∴ Fourier transform of a continuous time aperiodic signals

Synthesis Equation; Inverse Transform:

$$x(t) = \int_{-\infty}^{\infty} x(f) \, e^{-j \, 2\pi \, ft} \, df \qquad ....(2.6)$$

Analysis Equation; Direct Transform:

$$x(f) = \int_{-\infty}^{\infty} x(t) \, e^{-j \, 2\pi \, ft} \, dt \qquad ....(2.7)$$

Similarly,

Fnergy of a aperiodic signals is given as,

$$E_x = \int_{-\infty}^{\infty} | x(t)|^2 \, dt \qquad ....(2.8)$$

2.5.3 The Fourier Series for Discrete-Time periodic signals:

Synthesis Equation:

$$x(n) = \sum_{k=0}^{N-1} C_k \, e^{j2\pi \, kn/N} \qquad ....(2.9)$$

Analysis Equation:

$$C_k = \frac{1}{N} \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N} \qquad \text{.........(2.10)}$$

The average power of discrete time periodic signal is given as

$$P_x = \frac{1}{N} \sum_{n=0}^{N-1} |x\,(n)|^2 \qquad \text{...........(2.11)}$$

2.5.4 The Fourier Transform for Discrete Time Aperiodic Signals:

Synthesis Equation, Inverse Transform

$$x(n) = \frac{1}{2\pi} \int_{2\pi} x\,(w)\; e^{jwn}\, dw \qquad \text{...........(2.12)}$$

Analysis Equation, Direct Transform:

$$x(w) = \sum_{n=-\infty}^{\infty} x(n)\; e^{-jwn} \qquad \text{.............(2.13)}$$

$\therefore$ The energy of a discrete time signal x(n) is given as,

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2 \qquad \text{.............(2.14)}$$

2.6 DFT - THE DISCRETE FOURIER TRANSFORM:

DFT plays an important role in many applications of digital signal processing, including Linear filtering, correlation analysis and spectrum analysis. Efficient algorithms are used for computation of DFT.

DFT is a set of N samples {x(k)} of the Fourier Transform x(w) for a finite duration sequence {x(n)} of length L ≤ N. The sampling of x(w) occurs at the N equally spaced frequencies wk = 2π k/N, k = 0, 1, 2....N. Hence, N point DFT of a finite duration sequence x(n) of length L ≤ N is defined as

$$x(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi\, kn/N} \qquad ....(2.15)$$

$$k = 0, 1 \ldots\ldots N-1$$

and the IDFT is,

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(k)\, e^{j2\pi\, kn/N} \qquad ....(2.16)$$

$$n = 0, 1 \ldots\ldots N-1$$

2.6.1 Direct Computation of the DFT:

The direct computation of DFT requires,

1. $2N^2$ evaluation of trigonometric function;
2. $4N^2$ real multiplications;
3. 4N (N-1) real additions; and
4. A number of indexing and addressing operations.

So, the development of computationally efficient algorithms for the DFT is made possible if we adopt a divide and conquer approach. This approach is based on the decomposition of an N point DFT into successively smaller DFTs. This basic approach leads to a family of computationally efficient algorithms known as FFT algorithms.

The divide and conquer approach is used to derive fast algorithms when the size of the DFT is restricted to be a power of 2 or a power of 4.

REFERENCES

1. "Introduction to Digital Signal Processing" by John G. Proakis, Dimitris G. Manolakis.
2. "Microprocessors And Interfacing" by Douglas Hall. McGraw Hill Publication. 1990
3. "Fundamentals of Digital Signal Processing" by L.C. Ludeman. Harper & Row Publishers, New York, 1986.
4. "Communication Systems" by B.P. Lathi. Willey, Eastern United, New Delhi, Banglore, Bombay.
5. "Signal Processing Algorithms" by Samuel D. Stearns, Ruth A. David. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.